A mixed-mode follow-up of panel refusers in the Dutch LFS

Reinder Banning and Barry Schouten

Discussion paper (09035)

Statistics Netherlands

The Hague/Heerlen, 2009

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2007–2008	= 2007 to 2008 inclusive
2007/2008	= average of 2007 up to and including 2008
2007/'08	= crop year, financial year, school year etc. beginning in 2007 and ending in 2008
2005/'06–2007/'08	= crop year, financial year, etc. 2005/'06 to 2007/'08 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

ISSN: 1572-0314

6008309035 X-10

A mixed-mode follow-up of panel refusers in the Dutch LFS

Reinder Banning and Barry Schouten

Summary: The Dutch Labour Force Survey (LFS) is a rotating panel survey, of which the first wave is CAPI and all the other waves are CATI. LFS statistics based on the first CAPI wave, especially employment status, show small but systematic differences with statistics based on the CATI waves.

In the past, various attempts at finding the cause of these differences have been made. Unfortunatley, they did not provide a clear answer. It is believed that the observed differences are the consequence of a mixture of effects (non-response, panel, reference period, mode) that cannot easily be disentangled.

In May-July 2005, all households that objected to participation in waves 2 to 5 of the LFS were re-approached using two mixed-mode strategies. In doing so we aimed to investigate the selection effects due to non-response. The first strategy was a mix of CATI and CAPI, where CAPI was assigned to households for which no phone number was available. The second strategy was a mix of CATI, web and paper, where web and paper were offered to households without a phone number.

In this paper, we discuss the two approaches. Although sample sizes are modest, we find indications that refusing households have more volatile employment behavior than regular panel households.

Keywords: Non-response, LFS, Mixed-mode design

1 Introduction

The Dutch *Labour Force Survey* (LFS) is a rotating household panel survey comprised of five waves of interviews. At the end of the first wave *Computer Assisted Personal Interviewing* (CAPI) interview, each household is asked whether or not it is willing to participate in four additional *Computer Assisted Telephone Interviewing* (CATI) interviews. Not every household reacts positively to this request although, typically, between 80% and 90% of the respondents do. Subsequently, the respondents are asked to provide a telephone number and are interviewed with time lags of three months. In 2005, a so-called follow-up effort was undertaken for those households which refused to take part in the regular LFS panel of April to July. The results of this follow-up effort are the topic of research in this paper.

The target population of the Dutch LFS consists of all inhabitants of the Netherlands of 15 years old and older, except for people living in institutions. Its main objective is to produce a set of statistics about the employment status of persons and households. Most statistics concern the population of '$15-64$' years. However, the LFS also produces statistics about persons of 65 years and older.

From the group of, so-called, LFS panel refusers, a selection has been made. The selected households have been re-approached in a final attempt to try and entice them to cooperate in the follow up panel survey. The reason for spending time and effort on trying to include these panel refusers in the survey lies in the fact that they participated in the original LFS three months previously. This means that we have an unprecedented wealth of information on the households from the group of non-respondents of the panel. This information is not only essential to the investigation of the response behavior of the re-approached panel refusers. It can also be exploited to tackle the issue whether the behavior in the labour market of the LFS panel refusers is different from that of the regular LFS respondents, or not.

1.1 Background information

The sampling frame of the LFS is the Dutch municipal population administration (translation: Gemeentelijke Basisadminstratie or GBA). The survey is a two-stage sample, where geographical stratification is based on the so-called COROP-classification. In the first stage clusters are formed by municipalities. From the clusters simple random samples without replacement are drawn consisting of addresses. The first-order inclusion probabilities differ only for age. Addresses with all inhabitants older than 64 years have a lower inclusion probability. Also in the allocation of addresses the sample is reduced for some interviewer districts due to workload or staffing of interviewers. Up to

four households can be interviewed per address, and within each household the maximum number of persons to be interviewed is set to eight. The selected households are interviewed face-to-face in *Computer Assisted Personal Interviewing* (CAPI). Proxy interviewing is allowed under certain circumstances.

1.2 Formulating the research questions

In this paper, we want to analyze the response behavior of the households included in the re-approach effort. Furthermore, we are interested in the labour market behavior of regular LFS panel respondents versus that of responding, re-approached LFS panel refusers. In order to answer these questions, we formulate the following two research questions that will be looked into in greater detail in this study:

Research question 1:

What are the demographic, geographic, and socioeconomic characteristics that distinguish a responding, re-approached LFS panel refuser from a non-responding one?

Research question 2:

What are the demographic, geographic, and socioeconomic characteristics that distinguish a responding, regular LFS panel participant from a responding, re-approached LFS panel refuser?

With the answers to these questions, we can deduce whether or not we can ignore the panel non-response, i.e. whether or not we can adjust for the non-response by means of a weighting model. Also, but less importantly, we can derive whether or not a follow-up is an effective instrument for getting 'new' respondents.

1.3 Details on the re-approach effort

The group of households participating in the re-approach effort is special in the sense that at an earlier point in time they refused to take part in the rotating LFS panel. Consequently, the manner in which these households are interviewed now, is different from the manner in which they were interviewed originally. An important difference between these two ways of taking an interview is, that only one individual per household is interviewed in the re-approach effort (by proxy if necessary), whereas the entire eligible household was interviewed in the regular LFS panel. Furthermore, the individual to be interviewed in the re-approach effort is the next person to celebrate his or her birthday (with respect to the date of the interview), see Cobben and Schouten (2007b).

In the re-approach effort, the households have been interviewed using different interviewing modes. Each of these modi operandi corresponds to a different interviewing technique. The four different interviewing techniques in question are listed below.

1. *Call Back Approach Computer Assisted Telephone Interviewing* (CBA-CATI)
2. *Call Back Approach Computer Assisted Personal Interviewing* (CBA-CAPI)
3. *Basic Question Approach Computer Assisted Telephone Interviewing* (BQA-CATI)
4. *Basic Question Approach Mail/Web Survey*(BQA-M/W)

The group of panel refusing households that is to be re-approached, has been randomly divided into two groups of equal size. One group will be submitted to re-approach efforts using one of the two *Call Back Approach* (CBA) techniques[1]. The other group will be re-approached by interviewers using either of the two *Basic Question Approach* (BQA) techniques[2]. Using one of the BQA interviewing techniques, the survey is limited to a number of key questions[3]. The main goal here is to obtain insight into any differences that might exist between respondents and non-respondents.

The choice between using the CBA-CATI or the CBA-CAPI technique for a particular (selected) household depends on the presence of a listed land-line telephone connection. The same criterion is used for deciding between employing the BQA-CATI or the BQA-M/W technique. As in the regular panel survey, proxy interviewing is allowed.

2 Outlining the approach to the research questions

2.1 The approach to research question 1

The first research question focusses on individuals (a household is a *population unit* and consists of individuals or household members; a household member is a *population element*) who have refused to take part in the LFS panel, but that

[1] Interviewing techniques 1 and 2 together, are referred to as the CBA interviewing techniques, or the CBA regime.

[2] Interviewing techniques 3 and 4 together are referred to as the BQA interviewing techniques, or the BQA regime.

[3] This approach was first proposed in Kersten and Bethlehem (1984). Observations revealed that persons refusing to participate in a survey, could often be persuaded to just answer a few choice questions. These choice questions came to be known as the *Basic Questions* of the survey.

have been approached again. The re-approach has, unfortunately, not been completely successful. Some individuals did respond to the re-approach effort, others did not. Traditionally, the group of non-respondents is the group we lack information about. However, the current situation is different because both the respondents and the non-respondents of the re-approach are part of the first wave LFS response which took place three months earlier.

For each individual or population element under consideration, we identify the random *response variable* R. Value 1 is assigned to this variable if the person concerned participates in the survey. In all other cases, value 0 is assigned to R. Distinction between individuals is made by means of subscript i.

$$R_i = \begin{cases} 1 & \text{response} \\ 0 & \text{non-response} \end{cases} . \tag{1}$$

This random response variable R_i is a binary variable, and can be modelled by means of a *Bernoulli trial*. Our intention is to construct a so-called *Generalized linear model* (GLM)[4] for response variable R_i, see Agresti (2002). This model will consist solely of variables that have been made available through the first wave LFS response. The GLM derived in this manner, is an explanatory model for the response variable in terms of the background information provided by the first wave LFS response. As such, it will shed light on the demographic, geographic, and socioeconomic make-up of each household member who participates in the re-approach to the LFS follow up panel survey.

In the re-approach effort, only one individual per household has been interviewed. More precisely, one population element has been selected to be interviewed for every re-approached population unit. The model parameters of the GLM fitted to the measurement data, are influenced by this selection procedure. It is possible to compensate for these effects, provided detailed knowledge of the selection procedure is available. The preferred format of the knowledge required is the sampling scheme.

For the purpose of formulating a sampling scheme that suitably describes the selection procedure employed in the re-approach effort, we assume that a population unit is internally homogeneous with respect to the response variable (i.e. the response behavior of the household as a whole is coincident with the response behavior of the interviewed household member). As a matter of consequence, the households into which the individuals have been divided can be interpreted as internally homogeneous strata (with respect to the response variable).

[4] GLMs extend ordinary regression models to encompass nonnormal response distributions and modeling functions of the mean. Three components specify a GLM: A random component identifies the response variable Y and its probability distribuiton; a systematic component specifies explanatory variables used in a linear predictor function; and a link function specifies the function of $E(Y)$ that the model equates to the systematic component, Agresti (2002).

The manner in which a re-approached individual has been selected from the household, may be viewed as a process of drawing a *simple random sample* (SRS) of size 1 from the group of eligible household members. Because of this point of view and the fact that one population element (i.e. a person) is interviewed per population unit (i.e. a household), we may assume the overall selection scheme to be adequately approximated by a *stratified random sampling* scheme, with an SRS sample of size 1.

Let *inclusion probability* π_{hk} denote the probability that individual k of household h is selected to be interviewed in the re-approach effort (i.e. the probability of selecting element k from population unit h into the sample). It is then clear that π_{hk} is constant within a unit element, as should be under the conditions of the SRS. However, the probability is allowed to change for different values of h (i.e. from household to household).

Let us assume that the number of strata (i.e. the number of households) is denoted by H, and that the size (i.e. the number of eligible individuals in a household) of stratum h is denoted by N_h $(h = 1, \cdots, H)$. The overall size of the data set as denoted by symbol N, can then be determined according to the formula:

$$N = \sum_{h=1}^{H} N_h \tag{2}$$

Subsequently, the inclusion probability for element k in stratum h may be evaluated as

$$\pi_{hk} = \frac{1}{N_h} \qquad \forall h = 1, \cdots, H \quad \wedge \quad k = 1, \cdots, N_h \qquad . \tag{3}$$

Obviously, this probability is constant within a stratum, as expected. With this formula for the inclusion probability, the sampling scheme is complete.

For analysis of the question whether there are differences between a responding re-approached panel refuser and a non-responding re-approached panel refuser or not, it might be worthwhile to pay special attention to the effects of the type of interviewing technique used (i.e. CBA versus BQA). In the current analysis, three different case situations are therefore distinguished. The distinctions between these case situations are based on the types of interviewing technique used.

1. Case 0 situation: the outcome of the survey for all four modes;
2. Case I situation: the outcome of the survey for the modes 'CBA-CATI' and 'CBA-CAPI';
3. Case II situation: the outcome of the survey for the modes 'BQA-CATI' and 'BQA-M/W'.

2.2 The approach to research question 2

The second research question focusses on the respondents from the LFS panel survey and the respondents from the re-approach effort. These two groups are mutually exclusive. A respondent has been interviewed either in the course of the LFS panel survey or in the course of the re-approach effort.

In order to facilitate an analysis of this research question, these two groups have been merged into a new group. Subsequently, a random variable Z_i (see equation (4); subscript i identifies the record) is introduced for each population element in this group. Value 1 is assigned to variable Z_i if the individual concerned originates from the group that participated in the LFS panel survey; the value of variable Z_i is set to zero for all other population elements.

$$Z_i = \begin{cases} 1 & \text{respondent from the group of LFS panel participants} \\ 0 & \begin{cases} \text{respondent from the group of re-approached} \\ \text{LFS panel refusers} \end{cases} \end{cases} \tag{4}$$

Variable Z_i is a binary variable that can be modelled by means of a Bernoulli trial, and our aim is to develop a GLM for it. This model will rely entirely on the background information provided by the first wave LFS response.

In the regular LFS panel, all the eligible members of a responding household are interviewed. This is not so for the eligible members of a responding household from the re-approach effort. Indeed, in such a household the individual to be interviewed is selected from the group of eligible household members. Clearly, there is a fundamental difference between the manner by which a respondent from the regular panel survey has been selected, and the manner by which a responding re-approached panel refuser has been selected. This difference between the applied selection procedures has an effect on the values of the model parameters of the GLM fitted on the measurement data. These effects can be compensated for in the subsequent behavioral analysis, provided an appropriate overall sampling scheme has been formulated.

For the purpose of formulating this overall sampling scheme, we start off with the total group of households that responded (i.e. the responding households from the regular panel or the re-approach effort). As was the case under research question 1, the participating households are viewed as as many *strata* of eligible individuals. The assumption is that every stratum (i.e. household) is internally homogeneous with respect to response behavior. This means that within a stratum, the probability of responding (upon being interviewed) is the same for every eligible household member, see Cobben and Schouten (2007). It is furthermore true that an eligible individual from the group of responding panel participants and an eligible individual from the group of responding re-approached panel refusers have, generally speaking, different probabilities of actually being interviewed.

Let H denote the total number of strata (i.e. households) in the data set; in addition, let H_{pan} be the number of strata associated with the regular LFS panel. As a matter of consequence, the number of strata submitted to the re-approach effort, is equal to $H_{apr} = (H - H_{pan})$. Assume next, that the numbering of the strata is such that the first H_{pan} strata are coincident with the H_{pan} households from the regular LFS panel survey. We can then say that stratum h for $h = (H_{pan} + 1), \cdots, H$ is, originally, a panel refusing household that has been re-approached (successfully). Next, consider stratum h ($h = 1, \cdots, H$). Let the number of eligible individuals in this stratum be denoted by N_h. The overall number of eligible household members who are present in the households under consideration, notation N, can then be evaluated as:

$$N = \sum_{h=1}^{H} N_h = N_{pan} + N_{apr} \tag{5}$$

with

$$N_{pan} = \sum_{h=1}^{H_{pan}} N_h \qquad \wedge \qquad N_{apr} = \sum_{h=(H_{pan}+1)}^{H} N_h \qquad . \tag{6}$$

At an earlier stage in this discussion, we observed that in a responding household in the regular LFS panel survey, all eligible household members are interviewed. The implication of this observation is, that the inclusion probability π_{hk} for individual k in household h is equal to 1 for these household members. Mathematically formulated, we therefore have:

$$\pi_{hk} = 1 \qquad \forall h = 1, \cdots, H_{pan} \quad \wedge \quad k = 1, \cdots, N_h \qquad . \tag{7}$$

In the re-approach effort, however, a single individual per household has been selected for the interview. As has been argued in Section 2.1, this selection procedure is akin to the drawing of an SRS of size 1 from the group of eligible household members per household. These arguments still hold true in the current circumstances. As a result we arrive at the following inclusion probability π_{hk} for individual k in household h.

$$\pi_{hk} = \frac{1}{N_h} \qquad \forall h = (H_{pan} + 1), \cdots, H \quad \wedge \quad k = 1, \cdots, N_h \qquad . \tag{8}$$

When we consider the question whether there is a difference between a respondent to the regular LFS panel and a respondent to the re-approach effort or not, it might be worthwhile for us to take into consideration that different interviewing techniques have been used. Because of this, we have identified three case situations that might be relevant to research question 2. These case situations are listed below and will be discussed individually in the remainder of this report.

1. Case 0 situation: the combined data set of the respondents to the panel and the respondents to the re-approach effort, is considered.

2. Case I situation: the combined data set of the respondents to the panel, and the respondents to the re-approach effort who have been by interviewed under the CBA-regime, is considered.

3. Case II situation: the combined data set of the respondents to the panel, and the respondents to the re-approach effort who have been by interviewed under the BQA-regime, is considered.

It should be noted that the manner in which case situations 2 and 3 are dealt with is different from the manner in which case situation 1 is dealt with. The reason behind it is a difference in the underlying sampling schemes, see Cobben en Schouten (2007a); Cobben and Schouten (2007b). For example, consider an arbitrary, re-approached, eligible would-be respondent. The probability of actually being interviewed in either of the four modes, is for this household member not the same as the probability of being interviewed under one of the two CBA modes. Following similar reasoning, this household member's probability of being interviewed in either mode is not the same as its probability of being interviewed under one of the two BQA modes.

When we separate the households interviewed under the CBA-regime from the households interviewed under the BQA-regime as in the Case I situation and the Case II situation respectively, we have to allow for an inclusion probability that is halved in value. Put differently, the inclusion probability $\tilde{\pi}_{hk}$ for the re-approached individual k in household h in either the Case I or the Case II situation, is equal to:

$$\tilde{\pi}_{hk} = \frac{1}{2N_h} \quad \forall h = (H_{pan} + 1), \cdots, H \quad \wedge \quad k = 1, \cdots, N_h \tag{9}$$

3 Analyzing research question 1

3.1 Introduction

In the data set that has been prepared in advance for research question 1, see Banning (2008), each record represents an individual. This data set is comprised of 763 records. The information gathered in the re-approach has been augmented with information provided by the first wave LFS response. Also, the information concerning the interviewing mode employed, has been stored in the variable **Mode**.

In order to be able to analyze the response behavior of the re-approached LFS panel refusers, the response variable R_i has been introduced, equation (1). We intend to explain the behavior of R_i in terms of the interviewing mode used, the demographic, the geographic and the socioeconomic characteristics of the interviewee's. We propose, therefore, to develop a logistic regression model

designed to predict the response behavior of an arbitrary individual given the mode of interview and his or her demographic, socioeconomic, and geographic characteristics.

The required demographic, socioeconomic and geographic characteristics for each individual will be provided by variables from the first wave LFS response. Together with the variable **Mode**, they may be regarded as the *auxiliary variables*[5] with respect to response variable R_i.

The demographic properties of a person (i.e. a record from the data set) are represented by the variables[6] **Household type**, **Ethnicity** and **Age** respectively. The first variable conveys information about the composition of the household that the person in question is part of. The second variable contains information with regard to the ethnicity of the person. The third and last demographic variable reflects the person's age.

The socioeconomic characteristics of each person are represented by the variables **Employment status** and **Education**. The former variable contains information with regard to the status of that person's position in the labour market. The latter variable records the highest level of education received by a person.

For the geographic characteristic of a person, we rely on the degree of urbanization of this person's place of residence. This degree of urbanization is stored in the variable **Urbanization**.

3.2 A frequency analysis of the auxiliary variables

The frequency distribution of the auxiliary variable **Mode** is available for inspection in Table 1. From the figures displayed, we conclude that the aforementioned preparation procedure was of little or no effect, as far as the balance between the number of people interviewed under the CBA-regime (380) and the BQA-regime is concerned.

Table 1. Frequency distribution of the variable Mode under research question 1

	Frequency	*Percentage*
CBA-CATI	229	30.0
CBA-CAPI	151	19.8
BQA-CATI	230	30.1
BQA-M/W	153	20.1
Total	763	100.0

[5] Definitions of the auxiliary variables are available in Appendix A.

[6] Note that these variables have been selected because they were deemed suitable for the purpose of constructing a model.

Demographic variables The demographic properties of a record are represented by the variables **Household type**, **Ethnicity** and **Age**. For all three case situations, the frequency distributions of these variables are contained in Table 26 in Appendix B (page 43). Comparisons between case situations are conveniently performed if we adopt a, so-called, *reference* situation. For the sake of convenience of argument, we select the case 0 situation as our reference situation throughout this subsection.

Looking closely at the frequency distribution of variable **Household type**, we notice that it does not change much over the different case situations. Indeed, absolute variation from the reference situation is less than or equal to 0.9%. The same type of conclusion holds true for variable **Ethnicity**. The level of variation (from the reference situation) reaches a maximum of only $\pm 2.0\%$ for the category 'First generation', according to Table 26.

The message from the frequency distribution of variable **Age** is a mixed message. Although the figures for the different case situations are similar, the level of variation from the reference situation we have to allow for is more pronounced at $\pm 3.8\%$. ■

Socioeconomic variables The variables under consideration are the variables **Employment status** and **Education level**. Their frequency distributions for all three case situations are given in Table 27 in Appendix B (page 44).

Variable **Employment status** deals with the issue whether members of a household form an active part of the labour force or not. On the basis of its frequency distribution, we conclude that the number of individuals taking part in the active labour force is almost in balance with the number of individuals taking part in the non labour force in the case 0 and the case I situation. In the case II situation, the difference between the two categories is 6%. Having said that, the variation per category over the three case situations amounts to $\pm 2.7\%$ from the reference situation.

The variable **Education level** contains information on the highest level of education received by the individual, see Table 27. The range of parameter variations over the three case situation is limited to $\pm 1.4\%$ with respect to the reference situation. The percentage of (interviewed) people that has enjoyed the education level corresponding to the category 'Low', is equal to 44.4%. This is the highest percentage in the table. ■

Geographic variable The variable **Urbanization** reflects the degree of urbanization of an individual's place of residence. Details of its frequency distribution are available in Table 28 (page 44). Per category, the variation of this variable over the three case situations from the reference situation, is equal to or smaller than $\pm 1.1\%$. ■

Summarizing, the population under investigation in each of the three case situations is, to all intents and purposes, invariant with respect to the demographic, the socioeconomic and the geographic variables.

3.3 On the levels of association between the response variable and the auxiliary variables

In this investigation that is preparatory to the logistic model building effort, we will test the levels of association between the response variable R_i and the auxiliary variables. In Section 2.1, three case situations have been introduced. Each of these case situations will be reviewed separately.

The frequency distributions of R_i for the case 0, I and II situations, are grouped together in Table 2. In the case 0 situation, the complete data set of the re-approach effort (763 records) is considered. The level of response under the full re-approach effort equals 60.9%. By selecting the results of the CBA-CATI and the CBA-CAPI interviewing techniques (i.e. the case I situation), the data set reduces from 763 records to 380 records. A response of 76.6% is now achieved. Clearly, the combined CBA interviewing techniques elicit a considerably higher response rate than the re-approach effort as a whole, for which the response rate stands at 60.9%.

Finally, by selecting the results of the BQA-CATI and the BQA-M/W techniques (i.e. the case II situation), the data set reduces from 763 records to 383 records. The response has dropped to 45.4% from the original 60.9% in the case 0 situation.

For the purpose of gauging the levels of association between R_i and each of the auxiliary variables, we employ the design-based Wald test, Koch et al (1975), and the adjusted Pearson's χ^2-test, Rao and Scott (1984). Both association tests are designed to test the (null-)hypothesis of independence H_0. This null-hypothesis of independence H_0 (see the formulation given below) will be rejected if the corresponding p-value lies below or is equal to an a priori selected threshold probability value $p_{threshold}$.

$$\{H_0 \quad : \quad \text{the stochastic response variable } R_i \text{ and the stochastic auxiliary variable concerned, are mutually independent}\} \quad .$$

In this study we have selected a threshold probability of 10%, i.e. $p_{threshold} = 0.10$. The reason for selecting this relatively high value for the threshold probability, lies in the explorative nature of this study. We attempt to build logistic regression models composed of variables whose contributions are significant. The testing of a hypothesis is, therefore, performed for the sole benefit of this model building effort.

For each covariate, the findings of the two aforementioned tests have been made available in Table 3. This table includes the calculated p-values for each test.

The results are presented in order of increasing p-value as determined in the Wald test.

Table 2. Frequency distribution of variable R_i for the re-approached LFS panel refusers

Case situation		*Frequency*	*Percentage*
Case 0	Response	465	60.9
	Non-response	298	39.1
	Total	763	100.0
Case I	Response	291	76.6
	Non-response	89	23.4
	Total	380	100.0
Case II	Response	174	45.4
	Non-response	209	54.6
	Total	383	100.0

Case 0 situation The conclusion we draw from the contents of Table 3 has two aspects. Its first and foremost aspect is, that the calculated p-values signify the rejectability of the null-hypothesis of independence for the variables **Mode** and **Age** according to both tests. Indeed, it can be argued that each of the levels of association between the response variable and the auxiliary variables, is strong.

The second aspect of the conclusion is, that the null-hypothesis of independence cannot be rejected for either of the variables **Urbanization**, **Employment status**, **Education level**, **Household type** and **Ethnicity**. All the calculated p-values simply exceed the threshold p-value. ■

Case I situation The conclusion we draw from Table 3 is characterized by two aspects. The most intruding aspect is the outcome stating that, according to the Wald test, the null-hypothesis is rejected for the variables **Ethnicity**, and **Mode**. According to the Pearson χ^2-test, however, the H_0 can be rejected for the variables **Ethnicity**, **Mode**, and **Urbanization**. Because both tests agree on the rejectability of H_0 for only two variables, we adopt the conservative point of view and accept the rejectability of the null-hypothesis for the variables **Ethnicity**, and **Mode**. For these two auxiliary variables there exist, arguably, strong individual levels of association with response variable R_i.

The second aspect of the conclusion states that H_0 cannot be rejected for the variables **Urbanization**, **Education level**, **Age**, **Household type** and **Employment status**. For this group of variables, it cannot be disproved that either one of them is independent from response variable R_i. ■

Case II situation The results of the two association tests have been presented in Table 3. From its contents, we draw a conclusion that is characterized by two aspects.

According to aspect no. 1 of the conclusion, both tests agree on the fact that we can reject H_0 for the variables **Mode**, **Ethnicity**, **Household type**, and **Age**. For these variables, the individual levels of association with response variable R_i are strong.

The second aspect of the conclusion states that the null-hypothesis of independence H_0 cannot be rejected for the variables **Urbanization**, **Education level** and **Employment status**. In other words, we cannot disprove that response variable R_i is independent from these particular auxiliary variables. ■

Table 3. Association tests between response variable R_i and each auxiliary variable in the different case situation

Variable	Wald test			Pearson's χ^2 test		
	F	df	p-value	χ^2	df	p-value
Case 0						
Mode	39.35	3	0.0000	40.77	3	0.0000
Age	3.30	2	0.0375	2.58	2	0.0781
Urbanization	1.55	4	0.1856	1.43	4	0.2226
Employment status	0.92	2	0.3974	0.87	2	0.4184
Education level	0.99	3	0.3978	1.09	3	0.3534
Household type	0.83	3	0.4763	0.63	3	0.5865
Ethnicity	0.38	2	0.6829	0.38	2	0.6857
Case I						
Ethnicity	10.59	2	0.0000	7.79	2	0.0005
Mode	10.60	1	0.0012	9.92	1	0.0018
Urbanization	1.83	4	0.1230	2.02	4	0.0911
Age	1.16	2	0.3149	1.12	2	0.3029
Education level	1.62	3	0.3227	1.03	3	0.3249
Household type	1.05	3	0.3717	0.89	3	0.4403
Employment status	0.93	2	0.3943	1.02	2	0.3606
Case II						
Mode	66.11	1	0.0000	58.60	1	0.0000
Ethnicity	6.45	2	0.0018	6.07	2	0.0024
Household type	2.97	3	0.0319	2.49	3	0.0638
Age	3.10	2	0.0462	2.78	2	0.0650
Urbanization	1.82	4	0.1249	1.69	4	0.1495
Education level	0.87	3	0.4571	0.75	3	0.5245
Employment status	0.54	2	0.5859	0.60	2	0.5479

3.4 Logistic regression modelling

Next, we construct a logistic regression model in which the response variable acts as the dependent variable and the 7 auxiliary variables act as the covariates. Each of the different case situations is considered in a separate analysis and yields its own model. These models will be developed on the basis of a complex survey design. Such a model is, therefore, not fitted by maximum likelihood. Instead, the method employed fits a generalized linear model to data from a complex survey design, with inverse-probability weighting and design-based standard errors.

For a given model, the effect of each covariate can be determined through application of a Wald test for the hypothesis that all coefficients associated with a particular regression term are zero (or have some other specified values). Once this Wald test has been evaluated for every covariate in the model, it is possible to determine and select which covariates are significant to the model as a whole and which covariates are not, purely on the basis of the calculated p-values. By using this selection procedure, we are able to obtain a logistic regression model with a minimal number of significant covariates.

The (initial) logistic regression model we start the model building effort off with, is comprised of the main effects of all covariates under consideration in this analysis, i.e.

$$\begin{aligned} R_i \quad \propto \quad & \textbf{Mode} + \textbf{Urbanization} + \textbf{Age} + \textbf{Household type} + \\ & \textbf{Education level} + \textbf{Employment status} + \\ & \textbf{Ethnicity} \end{aligned} \quad . \tag{10}$$

This model is fitted on the observed measurement data for each case situation. Subsequently, the aforementioned Wald test is evaluated for each of the model covariates. The collective outcome of these tests is gathered by case situation and presented in Table 4. In this table, the covariates are presented in order of increasing p-value.

Case 0 situation According to the contents of Table 4, there is only one covariate for which the evaluated $p-$value lies below or is equal to $p_{threshold}$. The covariate in question is the variable **Mode**. This means that in initial model (10), none of the three demographic variables and none of the two socioeconomic variables is significant when it comes to explaining response behavior. Note, nevertheless, that the calculated p-value for the covariate **Urbanization** does not differ much from $p_{threshold}$. It is for this reason that in our first attempt at building a reduced order model, we will (logistically) regress the covariates **Mode**, and **Urbanization** on the response variable R_i. This means that a logistic regression model of the form as given next, is fitted on the response data.

$$R_i \quad \propto \quad \textbf{Mode} + \textbf{Urbanization} \quad . \tag{11}$$

Although initial logistic regression model (10) does not contain any cross products of the covariates (it only contains their main effects), the indirect effects of possible relationships between the covariates on the response behavior, did contribute to the calculated $p-$values (given in Table 4). By reducing the number of covariates in the model, we have reduced the amount of indirect effects that could potentially influence the model parameters. It is for this reason that we will apply the Wald test included in the model selection procedure to fitted model (11), just to make sure that all covariates in the model are indeed significant.

Table 4. Goodness of fit results for model (10) in the different case situations

Case situation	*Covariate*	χ^2	df	p-value
Case 0	**Mode**	98.28	3	0.0000
	Urbanization	7.23	4	0.1241
	Age	2.45	2	0.2944
	Household type	2.82	3	0.4202
	Education level	1.64	3	0.6509
	Ethnicity	0.62	2	0.7344
	Employment status	0.01	2	0.9933
Case I	**Education level**	295.01	3	0.0000
	Ethnicity	7.90	2	0.0193
	Mode	3.09	1	0.0786
	Age	1.93	2	0.3803
	Urbanization	2.49	4	0.6462
	Employment status	0.25	2	0.8818
	Household type	0.42	2	0.9371
Case II	**Mode**	36.99	1	0.0000
	Household type	6.36	3	0.0952
	Age	2.81	2	0.2459
	Urbanization	5.35	4	0.2530
	Ethnicity	2.03	2	0.3621
	Employment status	0.66	2	0.7202
	Education level	0.95	3	0.8132

The results of the test are presented in Table 5. They demonstrate that covariate **Mode** is significant and that covariate **Urbanization** is not, confirming the result in Table 4. As a matter of cause, the variable **Urbanization** is permanently removed from the model.

The final, reduced, linear logistic regression model to be considered, is the model:

$$R_i \quad \propto \quad \textbf{Mode} \quad . \tag{12}$$

Table 5. Goodness of fit results for model (11)

Covariate	χ^2	df	*p*-value
Mode	105.33	3	0.0000
Urbanization	6.68	4	0.1537

The evaluated model is submitted to the Wald test previously mentioned, in order to verify its validity. The results in Table 6 clearly demonstrate that the covariate **Mode** is significant.

Table 6. Goodness of fit results for model (12)

Covariate	χ^2	df	*p*-value
Mode	103.76	3	0.0000

The reference point of model (12) is defined by the category 'CBA-CATI' of covariate **Mode**. The point represents a re-approached LFS panel refuser who has been interviewed using the CBA-CATI interviewing technique. Through the process of fitting this model on the measurement data, we estimate the probability of encountering a responding, re-approached panel refuser on the condition of knowing the interviewing technique used. Formulated mathematically, the probability estimated is the conditional probability:

$$\Pr(R_i = 1 \quad | \quad \textbf{Mode}) \quad .$$

The results of the model fitting exercise are presented in Table 7. It contains the parameter values, the standard error values and the calculated *p*-values for all covariates concerned. The reference category for covariate **Mode** represents the category 'CBA-CATI'.

Table 7. The fitted parameters of logistic regression model (12)

Covariates	*Parameter*	*St error*	*t-value*	*p-value*
Intercept	0.8723	0.1630	5.350	0.0000
Mode				
CBA-CAPI	0.9018	0.2916	3.092	0.0021
BQA-CATI	−0.2790	0.2227	−1.253	0.2106
BQA-M/W	−2.2160	0.2736	−8.100	0.0000

From the values of the model parameters we deduce, that the probability of encountering a responding, re-approached panel refuser increases considerably with respect to the reference probability, when we change the interviewing technique from category 'CBA-CATI' to category 'CBA-CAPI'. Moreover, this change in interviewing technique is a significant change. The aforementioned probability decreases when we alter the interviewing technique from category

'CBA-CATI' to category 'BQA-CATI'. This change in interviewing technique is, however, not considered to be significant. Finally, a change in interviewing technique from category 'CBA-CATI' to category 'BQA-M/W' results in an even smaller conditional probability of meeting a responding, re-approached panel refuser. This time around, the change in interviewing technique is characterized as being significant.

In conclusion, when the results of the re-approach effort for all interviewing techniques are considered, only covariate **Mode** turns out to be significant in predicting the response behavior. The interviewing modi 'CBA-CATI' and 'CBA-CAPI' provide the best chance of finding respondents. The CAPI interviewing technique produces the highest predicted response rate. ■

Case I situation When we look at the contents of Table 4, we cannot but notice that the value of the χ^2-criterion is unusually large for covariate **Education level**. In point of fact, it is at a value of 295.01 by far the highest χ^2-criterion value in the table. How can this be explained?

The frequency distribution of covariate **Education level** is available in Table 27. According to its contents, the distribution of variable **Education level** is severely skewed: a coverage percentage of 0.8% for one category against coverage percentages of 45.8%, 34.7%, and 18.7% for the three remaining categories. It is believed that the skew distribution of covariate **Education level** is responsible for the high value of the χ^2-criterion, what we consider to be some kind of numerical effect only.

We attempt to prevent this numerical effect from occurring by re-categorizing the covariate. The four categories of **Education level** are easily collapsed into three categories because the smallest category (category 'Missing') only contains 3 individuals. These three individuals are all transferred to the largest category, category 'Low'.

Relying on the re-defined covariate **Education level**, the measurement data are fitted on initial model (10). The significance of each covariate in the model is subsequently determined through application of the Wald test that has been introduced earlier. The results of these tests are presented below in Table 8. Upon comparison with Table 4, we find that the covariate **Education level** has changed from the most significant covariate to the least significant covariate. Fortunately, the effects of the new covariate on the calculated χ^2-criterion value and p-value of the other covariates, are limited.

On the basis of the numbers in Table 8, we conclude that only covariates **Ethnicity**, and **Mode** are significant. The reduced logistic regression model to be fitted on the response data, will be of the following form:

$$R_i \quad \propto \quad \mathbf{Ethnicity} + \mathbf{Mode} \quad . \tag{13}$$

Table 8. Goodness of fit results for model (10) with the collapsed covariate

Case situation	*Covariate*	χ^2	df	p-value
Case I	**Ethnicity**	7.76	2	0.0206
	Mode	3.28	1	0.0701
	Age	1.92	2	0.3824
	Urbanization	2.67	4	0.6138
	Employment status	0.28	2	0.8663
	Household type	0.50	2	0.9184
	Education level	0.16	2	0.9253

For the purpose of evaluating the possible significance of each covariate in reduced model (13), we apply the Wald test mentioned previously. The results are presented in Table 9. The calculated p-values from this table tell us that the two covariates of reduced logistic regression model (13) are indeed significant.

Table 9. Goodness of fit results for model (13)

Covariate	χ^2	df	p-value
Ethnicity	9.33	2	0.0094
Mode	5.10	1	0.0239

The reference categories for the covariates **Ethnicity** and **Mode** have been selected as 'First generation non-native', and 'CBA-CATI' respectively[7]. The reference point of our model represents, therefore, an individual characterized by the categories 'First generation non-native' (covariate **Ethnicity**) and 'CBA-CAPI' (covariate **Mode**),

In model 13, we (logistically) regress **Ethnicity**, and **Mode** on response variable R_i. In doing so, we estimate the conditional probability of meeting a responding, re-approached panel refuser under the condition of knowing this person's ethnicity and knowing the used interviewing technique given that we only consider the results obtained by the two CBA interviewing techniques, i.e. we estimate the conditional probability:

$$\Pr\left(R_i = 1 \quad | \quad \textbf{Ethnicity} \wedge \textbf{Mode} \wedge (\textbf{Mode} = \text{'CBA-CATI'} \vee \textbf{Mode} = \text{'CBA-CAPI'})\right)$$

The results of the model fitting exercise are presented in Table 10. It is comprised of the actual model parameter values, the standard errors (on these model parameters), the corresponding t-values, and the calculated p-values.

Relative to the conditional probability associated with the reference point, the model predicts an increase in probability when a change is made from category

[7] The effects of these reference categories on the probability to be estimated, are all part of the 'Intercept' parameter.

Table 10. The fitted parameters of logistic regression model (13)

Covariates	*Parameter*	*St error*	*t-value*	*p-value*
Intercept	1.6050	0.3739	4.292	0.0000
Ethnicity				
Second generation	1.2952	1.1020	1.175	0.2406
Natives	−0.8797	0.3840	−2.291	0.0225
Mode				
CBA-CAPI	0.6683	0.2958	2.259	0.0244

'First generation non-native' to 'Second generation non-native' for covariate **Ethnicity**. Put differently, the model indicates that the probability of finding a responding, re-approached panel refuser of second generation ethnicity is bigger than the probability of finding a responding, re-approached panel refuser of first generation ethnicity. The considered change is, however, not significant.

The model parameters foretell a decrease in probability (with respect to the probability of the reference point) for the change in ethnicity from 'First generation non-native' to 'Natives'. Expressed in words, the probability of finding a responding, re-approached panel refuser of indigenous ethnicity is distinctly smaller than the probability of meeting a responding, re-approached panel refuser of first generation non-native ethnicity. This change in categories is strongly significant.

The second covariate under consideration is the covariate **Mode**. According to the calculated model parameter, the probability of encountering a responding, re-approached panel refuser increases when a change from category 'CBA-CATI' to category 'CBA-CAPI' is effectuated for the covariate **Mode**. In other words, the model predicts that the probability of finding a responding re-approached panel refuser that was interviewed by means of 'CBA-CAPI' is bigger than the probability of finding a responding re-approached panel refuser that has been interviewed by means of 'CBA-CATI'. Furthermore, the change in categories under consideration is significant.

In conclusion, the two modi 'CBA-CATI' and 'CBA-CAPI' are especially successful in obtaining a response from people with the ethniticity 'First generation non-native' and 'second generation non-native'. Of the two interviewing techniques considered, the CAPI technique yields the highest predicted response rate. ■

Case II situation Given the contents of Table 4, we find that the calculated p-value lies below the $p_{threshold}$ for the covariates **Mode** and **Household type**. The logistic regression model build to fit the response variable R_i depends,

therefore, on the two covariates **Mode** and **Household type**.

$$R_i \quad \propto \quad \textbf{Mode} + \textbf{Household type} \quad . \tag{14}$$

We can determine the validity of the fitted model through verification of the significance of each of the two covariates in (14). The significance of a covariate in a model, can be ascertained by means of a goodness of fit analysis (i.e. the Wald test included in the model selection procedure). The results of this analysis are presented in Table 11.

Table 11. Goodness of fit results for model (14)

Covariate	χ^2	df	p-value
Mode	48.84	1	0.0000
Household type	2.70	3	0.4395

According to the calculated p-values, the covariate **Household type** is not significant; only covariate **Mode** is a significant. We propose, therefore, to fit the measurement data on the new, reduced linear logistic regression model of the form:

$$R_i \quad \propto \quad \textbf{Mode} \quad . \tag{15}$$

In order to test the validity of this model, we submit the results of the model fitting effort to the Wald test that has been used previously for such a purpose. The outcome of the test in Table 12 tells us that covariate Mode is a significant covariate in this particular model.

Table 12. Goodness of fit results for model (15)

Covariate	χ^2	df	p-value
Mode	52.56	1	0.0000

In a model with a single covariate, the reference point is readily identified. In this particular instance, the reference category is the category 'BQA-CATI' of the covariate **Mode**. As a result, the reference point represents a re-approached individual who participated in a CATI interview under the BQA-regime.

By fitting this model on the measurement data, we effectively estimate the conditional probability of meeting a responding panel refuser under the condition of possessing knowledge about the interviewing technique that has been employed given that either one of the two interviewing techniques under the BQA-regime has been used. In mathematical terms, we estimate the probability:

$$\Pr(\, R_i = 1 \quad | \quad \textbf{Mode} \wedge (\textbf{Mode} = \text{'BQA-CATI'} \vee \textbf{Mode} = \text{'BQA-M/W'})) \tag{16}$$

The outcome of the model fitting effort is presented in Table 13. The figures in this table show that a decrease in probability (16) will be the effect of changing

the value of covariate **Mode** from category 'BQA-CATI' to category 'BQA-M/W'. Furthermore, this effect can be considered to be significant.

In conclusion, the interviewing modus 'BQA-M/W' is the least effective of the two BQA-interviewing modi, in obtaining a response from a re-approached LFS panel refuser. ■

Table 13. The parameters of logistic regression model (15)

Covariates	*Parameter*	*St error*	*t-value*	*p-value*
Intercept	0.5932	0.1518	3.907	0.0001
Mode				
BQA-M/W	−1.9370	0.2672	−7.250	0.0000

4 Analyzing research question 2

4.1 Introduction

The data set prepared for research question 2 consists of the records of responding individuals who partook in the panel survey, together with the records of individuals who responded in the re-approach effort. The former group is the largest in size, 16,863 people; the latter group is considerably smaller in size: 465 people. All in all, the data set contains 17,328 records.

In the analysis of research question 1, three demographic variables, two socioeconomic variables and one geographic variable have been selected from the original LFS response to provide background information. We will also exploit these variables as background information in our investigations into research question 2. In particular, the socioeconomic variable **Employment status** forms an essential part of the background information and refers to a person's position in the labour market at the time of the LFS survey.

The data set under consideration consists of respondents to either the panel survey or the re-approach effort. Both of these surveys contain a question regarding a person's current position in the labour market. This information is stored in the variable **Employment status 2**, and subsequently added to the list of socioeconomic variables available as background information. The reason for including variable **Employment status 2** in this list is, simply, that the inclusion provides knowledge regarding a person's position in the labour market at a second and consecutive time instant.

4.2 A frequency analysis of the auxiliary variables

According to the original figures in Table 1, the number of individuals interviewed in the re-approach effort by means of the CBA-CATI or the CBA-CAPI

interviewing technique (380 in total), is almost in balance with the number of individuals interviewed by means of the BQA-CATI or the BQA-M/W interviewing technique (383 in total). These numbers include both responding and non-responding individuals. The data set considered now is, however, different. Indeed, it only contains records of the responding individuals originating from the LFS panel response or the re-approach effort.

The make up of the data set[8] in terms of the type of interviewing techniques used, is made visible in Table 14. Clearly, the respondents to the regular LFS panel dominate the data set.

Table 14. Make-up of the data set with respect to the interviewing mode

	Frequency		*Percentage*	
CBA-CATI	164	(311)	1.0	(1.7)
CBA-CAPI	127	(224)	0.7	(1.3)
BQA-CATI	142	(295)	0.8	(1.7)
BQA-M/W	32	(54)	0.2	(0.3)
LFS panel	16,863	(16,863)	97.3	(95.0)
Total	17,328	(17,747)	100.0	(100.0)

Demographic variables The frequency distributions of the variables **Household type**, **Ethnicity** and **Age** are available in Table 29 (page 45). The contents of this table cover the three case situations as defined in section 4.1. We conclude that, as far as the variables under consideration are concerned, variation over the different case situations is limited. To be precise, category variations lie within ±0.4% from the reference situation, i.e. the case 0 situation.

When we compare the figures from Table 29 to those of Table 26, we notice some marked differences. Especially for the variable **Household type**. The category 'Couple +' has a far greater representation under research question 2 than it had under research question 1. The conclusion therefore is, that in terms of the demographic variables, we are dealing with distinctly different populations under the two research questions. ■

Socioeconomic variables The frequency distributions of variables **Employment status**, **Employment status 2** and **Education level** are available in Table 30 (page 46). As with the demographic variables discussed previously, the socioeconomic variables show little variation (±0.4% from the reference situation) over the three different case situations.

[8] In this table, the numbers in parentheses represent the weighted situation, i.e. the situation in which the numbers have been weighted with the inclusion probability.

A comparison[9] of Table 30 with Table 27 reveals some interesting discrepancies. It would seem that when we consider respondents only, we focus on people who are, on average, more often part of the active labour force and who are better educated, than when we consider respondents and non-respondents grouped together. In conclusion, under research question 2, the people considered are different from the people considered under research question 1 as far as these socioeconomic variables are concerned. ■

Geographic variable The frequency distribution of variable **Urbanization** is given in Table 31, (page 46). We conclude that the observed range of variation of $\pm 0.1\%$ from the reference situation, is small. Furthermore, the category 'Strong' is the largest in size. Upon comparison of the figures from this table to the corresponding frequency table under research question 1, we find that they are quite similar. ■

In summary, the different populations under investigation in the respective case situations are, for all practical purposes, invariant with respect to the demographic, the socioeconomic and the geographic variables.

4.3 On the levels of association between the output variable and the auxiliary variables

In this investigation we conduct a univariate association analysis between the variable Z_i on the one hand, and the different auxiliary variables on the other hand. Earlier, we introduced three case situations for research question 2, see Section 2.2. Here, each of these case situations will be investigated separately. The frequency distribution of variable Z_i under the different case situations is presented[10] in Table 15.

We will reject or not reject, the hypothesis of independence H_0 (see the formulation given below) on the basis of the outcome of the design-based Wald test (Koch et al (1975)) and the adjusted Pearson's χ^2-test (Rao and Scott (1984)).

$$\{H_0 \quad : \quad \text{the stochastic variable } Z_i \text{ and the specified stochastic auxiliary variable are mutually independent}\}$$

The results of the univariate association analysis are given in Table 16. The variables are presented from top to bottom, in order of increasing p-value from the Wald test.

[9] Note that the variable **Employment status 2** is not featured under research question 1. We can, therefore, only compare it to the variable **Employment status** under research question 2.

[10] In this table, the numbers in parentheses represent the weighted situation, i.e. the situation in which the numbers have been weighted with the inclusion probability.

Table 15. Frequency distribution of variable Z_i for the different case situations

	Frequency		*Percentage*	
Case 0				
Re-approach resp	465	(884)	2.7	(5.0)
Panel respondent	16,863	(16,863)	97.3	(95.0)
Total	17,328	(17,747)	100.0	(100.0)
Case I				
Re-approach resp -CBA-	291	(1,070)	1.7	(6.0)
Panel respondent	16,863	(16,863)	98.3	(94.0)
Total	17,154	(17,933)	100.0	(100.0)
Case II				
Re-approach resp -BQA-	174	(698)	1.0	(4.0)
Panel respondent	16,863	(16,863)	99.0	(96.0)
Total	17,037	(17,561)	100.0	(100.0)

Case 0 situation The results of the two tests are clear. All the calculated p-values yield, pair-wise, the same outcome. As a matter of consequence, we will refer to the results of the Wald test only. The conclusion to be drawn on the basis of the contents of Table 16 has two aspects.

Aspect no. 1 refers to the findings that the null-hypothesis of independence can be rejected for variables **Employment status 2**, **Employment status**, **Education level**, **Ethnicity**, **Age**, and **Household type**. In other words, (strong) levels of association exist between variable Z_i and each variable from this group of six variables.

The second aspect of the conclusion states that H_0 cannot be rejected for variable **Urbanization**. Consequently, the independence of variable Z_i and variable **Urbanization** cannot be disproved with sufficient significance. ■

Case I situation The results for this case situation display a pattern that is highly similar to the pattern observed in the data for the case 0 situation. For instance, the results of two tests agree with one another almost perfectly, see Table 16. This means that we can safely base our analysis on the outcome of the Wald test.

The first aspect of the conclusion states that null-hypothesis H_0 can be rejected for the variables **Employment status 2**, **Employment status**, **Education level**, **Ethnicity**, **Age**, and **Household type**. Once more, we come to the conclusion that there are strong levels of association between variable Z_i and these six variables.

Aspect no. 2 of the conclusion refers to the fact that H_0 is not rejected for covariate **Urbanization** in lieu of the calculated p-value being larger than the

threshold p-value. The hypothesis that variable Z_i is independent from variable **Urbanization** can, therefore, not be disproved. ■

Case II situation The results of the two tests for the case II situation, see Table 16, are less in agreement with one another than they were in the case 0 situation and/or the case I situation. These disagreements do, however, not run so deep that they influence our decision to reject or not reject the null-hypothesis of independence for any of the covariates under consideration.

According to either test, the null-hypothesis H_0 can be rejected for the variables **Employment status 2**, **Age**, **Employment status**, **Ethnicity** and **Education level**. The null-hypothesis is not rejected for the two remaining variables **Urbanization** and **Household type**. ■

Table 16. Association tests between variable Z_i and each auxiliary variable

	Wald test			Pearson's χ^2 test		
Variable	F	df	p-value	χ^2	df	p-value
Case 0						
Employment status 2	22.53	2	0.0000	26.98	2	0.0000
Employment status	19.10	2	0.0000	22.80	2	0.0000
Education level	12.25	3	0.0000	15.16	3	0.0000
Ethnicity	14.94	2	0.0000	32.32	2	0.0000
Age	14.83	2	0.0000	15.09	2	0.0000
Household type	3.08	3	0.0264	3.11	3	0.0290
Urbanization	0.06	4	0.9933	0.06	4	0.9936
Case I						
Employment status	18.20	2	0.0000	22.78	2	0.0000
Employment status 2	16.64	2	0.0000	23.52	2	0.0000
Education level	10.94	3	0.0000	14.30	3	0.0000
Ethnicity	13.87	2	0.0000	33.21	2	0.0000
Age	12.28	2	0.0000	13.59	2	0.0000
Household type	4.15	3	0.0060	4.48	3	0.0049
Urbanization	0.62	4	0.6453	0.60	4	0.6630
Case II						
Employment status 2	6.36	2	0.0017	6.95	2	0.0010
Age	5.44	2	0.0044	6.17	2	0.0024
Employment status	4.90	2	0.0075	4.46	2	0.0116
Ethnicity	3.02	2	0.0490	4.98	2	0.0069
Education level	2.32	3	0.0730	2.67	3	0.0468
Urbanization	0.43	4	0.7892	0.39	4	0.8149
Household type	0.22	3	0.8859	0.24	3	0.8511

4.4 Logistic regression modelling

The findings of the univariate association level analyses between variable Z_i and the covariates (see Section 4.3) state, that strong levels of association do exist between variable Z_i and each of the covariates **Employment status**, and **Employment status 2** in all three case situations. In a separate investigation, we have fitted the measurement data to a logistic regression model comprised of the main effects of the covariates (see below). The resulting model has subsequentely been validated.

$$\begin{aligned} Z_i \quad \propto \quad & \textbf{Urbanization} + \textbf{Age} + \textbf{Household type} + \\ & \textbf{Employment status} + \textbf{Employment status 2} + \\ & \textbf{Education level} + \textbf{Ethnicity} \end{aligned} \quad . \qquad (17)$$

The results of the Wald test revealed, that the covariates **Employment status**, **Employment status 2**, and **Urbanization** were not significant (the covariates **Ethnicity**, **Education level**, **Age** and **Household type** were all significant). This result prompted us to leave the main effects of these variables out of the model, and to introduce the cross product of the first two (of the former three) variables as a new covariate[11]. The rationale behind this is, that we know these two variables to be causally related.

The logistic regression model we, therefore, start off with is the model of the form:

$$\begin{aligned} Z_i \quad \propto \quad & \textbf{Urbanization} + \textbf{Age} + \textbf{Household type} + \\ & (\textbf{Employment status} \times \textbf{Employment status 2}) + \\ & \textbf{Education level} + \textbf{Ethnicity} \end{aligned} \quad . \qquad (18)$$

The measurement data is fitted to the initial logistic regression model under all of the different case situations. For each of the three fitted models, the Wald test from the model selection procedure[12] is evaluated, and the results are presented in Table 17. In this table, the covariates are recited in order of increasing p-value.

Case 0 situation The calculated p-values in Table 17 demonstrate that all covariates but one, are significant; the only not significant covariate is the covariate **Urbanization**. This means that all demographic and socioeconomic variables are significant covariates in logistic regression model (18).

The modelling effort is henceforth confined to the logistic regression of the covariates **Ethnicity**, **Education level**, **Age**, **Household type** and **Employment status**×**Employment status 2** on variable Z_i. Put differently, we try to fit the observed data on the logistic regression model formulated next:

[11] The frequency distribution of this composite covariate is available in Table 32, (page 47).

[12] The model selection procedure outlined in Section 3.4, employs a Wald test to determine the significance of a covariate (in the model). It relies on the calculated p-value for each covariate and can be repeatedly applied for different covariates.

$$Z_i \propto \textbf{Age} + \textbf{Household type} + \textbf{Education level} + (\textbf{Employment status} \times \textbf{Employment status 2}) + \textbf{Ethnicity} \quad . \tag{19}$$

The reduced model is subjected to the Wald test from the model selection procedure. The outcome of the test is made available in Table 18. According to the calculated p-values, all the covariates in the linear regression model are significant.

Table 17. Goodness of fit results for model (18) in the different case situations

Covariate	χ^2	df	p-value
Case 0			
Empl status×Empl status 2	178.13	8	0.0000
Ethnicity	50.30	2	0.0000
Education level	19.81	3	0.0000
Age	14.45	2	0.0000
Household type	9.99	3	0.0187
Urbanization	1.20	4	0.8776
Case I			
Empl status×Empl status 2	99.18	8	0.0000
Ethnicity	47.91	2	0.0000
Education level	15.66	3	0.0013
Household type	13.52	3	0.0036
Age	4.17	2	0.1241
Urbanization	0.18	4	0.9960
Case II			
Empl status×Empl status 2	16,690.38	8	0.0000
Age	15.00	2	0.0006
Ethnicity	10.01	2	0.0069
Education level	6.06	3	0.1089
Urbanization	2.35	4	0.6721
Household type	1.14	3	0.7650

Table 18. Goodness of fit results for reduced model (19)

Covariate	χ^2	df	p-value
Empl status×Empl status 2	178.66	8	0.0000
Ethnicity	50.85	2	0.0000
Education level	20.30	3	0.0001
Age	14.84	3	0.0006
Household type	9.27	2	0.0259

The reference point of model (19) represents a responding individual characterized by the reference categories. In turn, the reference categories are defined by the categories '$15-24$', 'Single', and 'First generation non-native' of the covariates **Age**, **Household type** and **Ethnicity**, respectively, and the category 'Low' of the covariate **Education level**. The reference point is complete with the category 'Active lf×Active lf' of the composite covariate (**Employment status**×**Employment status 2**).

When this logistic regression model is fitted on measurement data, we effectively estimate the probability of meeting a responding panel participant under condition of knowing the individual's age class, household make-up, education level, ethnicity, and history of labour force participation. Mathematically speaking, we estimate the following conditional probability:

$$\Pr\left(Z_i = 1 \quad | \quad \begin{array}{l} \textbf{Age} \wedge \textbf{Household type} \wedge \textbf{Education level} \wedge \\ (\textbf{Employment status} \times \textbf{Employment status 2}) \wedge \\ \textbf{Ethnicity} \end{array}\right) .$$

The results of the model fitting effort are presented in Table 19. This table contains the calculated model parameter values, their standard errors, the calculated p-values, and the corresponding t-values.

A predicted decrease (with respect to the conditional probability of the reference point) in the conditional probability of meeting a responding panel participant is equivalent to a predicted increase in the conditional probability of meeting a responding re-approached panel refuser.

The calculated p-values indicate, that all but five changes from the reference point are significant. The exceptions are the change from category 'Single' to category 'Single +' (covariate **Household type**), the change from category 'Low' to category 'Missing' (covariate **Education level**), the change from category 'Active lf×Active lf' to either of the three categories 'Inactive lf×Active lf', 'Inactive lf×Inactive lf', and 'Inactive lf × Non lf' (covariate (**Employment status**×**Employment status 2**)).

The predicted increase is quite large in the situations where a change from reference category 'Active lf×Active lf' to either one of the following categories: 'Active lf×Inactive lf', or 'Active lf×Non lf ', or 'Non lf×Active lf' is considered. In other words, the re-approach effort is quite effective in rooting out those individuals who have undergone a change in their position in the labour market.

It is worth noting that the model parameters for the covariate Age, are negative. For the remaining covariates, the parameter values are positive. This means that the re-approach effort is useful in finding responding participants aged 25 and older.

In conclusion, when the respondents from the re-approach effort are considered together with the regular panel respondents, then the re-approach effort is suc-

cessful in finding people who are aged 25 and older and whose position in the labour market has changed. ■

Table 19. The fitted parameters of logistic regression model (19)

Covariates	*Parameter*	*St error*	*t-value*	*p-value*
Intercept	2.2843	0.2558	8.931	0.0000
Age				
$25 \leq$ age ≤ 54	−0.3306	0.1894	−1.745	0.0809
$55 \leq$ age ≤ 64	−0.7947	0.2150	−3.696	0.0002
Household type				
Single +	0.1971	0.2108	0.935	0.3499
Couple	0.3509	0.1356	2.589	0.0096
Couple +	0.3519	0.1415	2.488	0.0129
Education level				
Average	0.4424	0.1276	3.466	0.0005
High	0.6172	0.1526	4.044	0.0000
Missing	0.1217	0.5898	0.206	0.8366
Ethnicity				
Second generation	0.6035	0.2389	2.526	0.0115
Natives	1.0476	0.1495	7.009	0.0000
Empl status×Empl status 2				
Act lf × Inact lf	−1.8182	0.5039	−3.608	0.0003
Act lf × Non lf	−2.4112	0.2259	−10.676	0.0000
Inact lf × Act lf	−0.8155	0.5194	−1.570	0.1164
Inact lf × Inact lf	0.0452	0.4402	0.103	0.9182
Inact lf × Non lf	−0.3693	0.5278	−0.700	0.4840
Non lf × Act lf	−2.4366	0.2404	−10.134	0.0000
Non lf × Inact lf	−0.8320	0.4268	−1.949	0.0513
Non lf × Non lf	−0.5400	0.1396	−3.868	0.0001

Case I situation The evaluated p-values in Table 17 indicate that, apart from covariates **Urbanization**, and **Age**, all covariates in model (21) are significant. Upon close inspection of the table we find that the calculated p-value for covariate **Age** of 12% does exceed the threshold value of 10%, but not by much. Because of this narrow margin, we maintain the presence of the covariate **Age** in our first attempt at model fitting. In other words, the first model fitting attempt is continued for the following reduced logistic regression model:

$$Z_i \propto \textbf{Household type} + \textbf{Ethnicity} + \textbf{Education level} + (\textbf{Employment status} \times \textbf{Employment status 2}) + \textbf{Age} \quad . \tag{20}$$

After the measurement data has been fitted on this (reduced) model and the actual model parameters have been determined, the validity of the model is tested. The outcome of the test is made available in Table 20. According to the calculated p-values, the covariate **Age** is not significant. The other covariates are significant.

Table 20. Goodness of fit results for reduced model (20)

Covariate	χ^2	df	p-value
Empl status×Empl status 2	99.36	8	0.0000
Ethnicity	52.10	2	0.0000
Education level	16.04	3	0.0011
Household type	14.11	3	0.0028
Age	4.21	2	0.1221

Adhering to the outcome of the validity test, the covariate **Age** is not included in the updated reduced model (see below) that will be considered for further use. Subsequently, the measurement data are fitted on model (21).

$$Z_i \propto \textbf{Household type} + \textbf{Ethnicity} + \textbf{Education level} + (\textbf{Employment status} \times \textbf{Employment status 2}) \quad . \tag{21}$$

The calculated model is also tested for its validity and the findings of the test are presented in Table 21. According to the calculated p-values, all the covariates of model (21) are significant.

Table 21. Goodness of fit results for reduced model (21)

Covariate	χ^2	df	p-value
Empl status×Empl status 2	126.36	8	0.0000
Ethnicity	50.35	2	0.0000
Household type	18.53	3	0.0003
Education level	16.06	3	0.0011

The reference point of model (21) is determined by the reference categories 'Single', and 'First generation non-native' of the covariates **Household type**, and **Ethnicity**, respectively, the reference category 'Low' of the covariate **Education level**, and the reference category 'Active lf×Active lf' of the composite covariate (**Employment status×Employment status 2**). By fitting this logistic regression model on the measurement data, we estimate the probability of meeting a responding panel participant under condition of knowing the individual's household make-up, education level, ethnicity, and history of labour force participation. Expressed in terms of mathematics, we estimate the

following conditional probability:

$$\Pr\left(Z_i = 1 \quad | \quad \begin{array}{l} \textbf{Household type} \wedge \textbf{Education level} \wedge \textbf{Ethnicity} \wedge \\ (\textbf{Employment status} \times \textbf{Employment status 2}) \wedge \\ (\textbf{Mode} = \text{'CBA-CATI'} \vee \textbf{Mode} = \text{'CBA-CAPI'}) \end{array}\right) .$$

The results of the model fitting effort are presented in Table 22. Looking only at the column of calculated p-values, we discover that a change from the reference point is significant for all but five covariate categories.

Of this last group of covariate categories, four are part of the covariate (**Employment status**×**Employment status 2**). Consequently, a change from the reference category is significant for the remaining four categories of this covariate. For the purpose of this analysis, we confine our argument to those categories for which a change from the reference value is significant.

All the model parameters are negative for the categories of the composite covariate (**Employment status**×**Employment status 2**). The fact that the model parameters are negative implies that the conditional probability of meeting a responding panel participant reduces when a change from reference value is contemplated for either one of these covariate categories. It has been pointed out before and will be emphasized here again, that this reduction in probability is equivalent to an increase in the conditional probability of meeting a responding, re-approached panel refuser. The absolute changes in probability are by far the largest for a change from reference category 'Active lf×Active lf' to one of the following three categories: 'Active lf×Non lf', 'Inactive lf×Active lf', and 'Non lf×Active lf'.

The conclusion based on the analysis of the fitted model (see Table 22) is, that the re-approach effort is efficient in finding respondents who have experienced a recent migration in the labour force market. ■

Case II situation In the case II situation, the results of the goodness of fit analysis, see Table 17, reveal that the value of the χ^2-criterion almost equals seventeen thousand for covariate (**Employment status**×**Employment status 2**). We suspect this extremely large criterion value to be a numerical effect caused by the peculiarities of the frequency distribution (see Table 32) of this particular covariate. For the purpose of verifying our suspicion we create a similar covariate with fewer categories.

The two variables **Employment status** and **Employment status 2** have three categories each. We propose to reduce the number of categories of each of these variables from three to two by collapsing the categories 'Inactive labour force' and 'Non labour force' into the (new) category 'Inactive/non labour force'. After crossing these two re-categorized variables, we obtain a covariate with four instead of nine categories. The frequency distribution of this new covariate is available in Table 33, (page 47).

Next, we repeat the exercise of fitting initial logistic regression model (18) onto the measurement data. Once more, the goodness of fit test is applied to the model parameters and the results are presented in Table 23. The results are much better and more in line with those for the Case 0 and the Case I situations, see Table 17. Actually, we see that the value of the χ^2-criterion has only changed dramatically (with respect to Table 17) for covariate (**Employment status**×**Employment status 2**). For the other covariates, the changes are limited and of little consequence.

Table 22. The fitted parameters of logistic regression model (21)

Covariates	*Parameter*	*St error*	*t-value*	*p-value*
Intercept	1.4424	0.2239	6.443	0.000
Household type				
Single +	0.4631	0.2324	1.993	0.0463
Couple	0.3716	0.1639	2.267	0.0234
Couple +	0.7171	0.1671	4.292	0.0000
Education level				
Average	0.4867	0.1584	3.072	0.0021
High	0.6571	0.1884	3.488	0.0005
Missing	−0.0980	0.7066	−0.139	0.8898
Ethnicity				
Second generation	0.5272	0.2717	1.940	0.0524
Natives	1.2220	0.1791	6.825	0.0000
Empl status×**Empl status 2**				
Act lf × Inact lf	−0.5970	0.7975	−0.749	0.4542
Act lf × Non lf	−2.3701	0.2727	−8.692	0.0000
Inact lf × Act lf	−1.2984	0.5470	−2.374	0.0176
Inact lf × Inact lf	−0.4596	0.4514	−1.018	0.3086
Inact lf × Non lf	−0.2650	0.8082	−0.328	0.7431
Non lf × Act lf	−2.3340	0.2774	−8.414	0.0000
Non lf × Inact lf	−0.5258	0.5174	−1.016	0.3095
Non lf × Non lf	−0.7479	0.1565	−4.779	0.0000

The contents of Table 23 signify that the presence of each of the covariates (**Employment status**×**Employment status 2**), **Age**, **Ethnicity** and **Education level** is significant. At the same time we find that the covariates **Urbanization** and **Household type** are not significant. As a matter of consequence, we proceed with the model fitting attempt using the following logistic regression model:

$$Z_i \;\propto\; \begin{array}{l} \mathbf{Age} + \mathbf{Education\ level} + \mathbf{Ethnicity} + \\ (\mathbf{Employment\ status} \times \mathbf{Employment\ status\ 2}) \end{array} . \qquad (22)$$

For the purpose of verifying the validity of the estimated model, the model is submitted to the previously mentioned Wald test. The ensuing results are recited in Table 24. According to the calculated p-values, all covariates considered are significant.

Table 23. Goodness of fit results for model (18) with a collapsed covariate

Covariate	χ^2	df	p-value
Empl status×Empl status 2	90.98	3	0.0000
Age	14.66	2	0.0007
Ethnicity	9.96	2	0.0069
Education level	5.90	3	0.1164
Urbanization	2.74	4	0.6031
Household type	1.10	3	0.7772

Table 24. Goodness of fit results for reduced model (22)

Covariate	χ^2	df	p-value
Empl status×Empl status 2	89.95	3	0.0000
Age	16.47	2	0.0003
Ethnicity	8.41	2	0.0149
Education level	6.36	3	0.0955

With model (22) explicitly known, we can evaluate estimates for the conditional probability of finding a responding panel participant under condition of said individual's age class, education level, ethnicity and history of labour force participation, given that only the results of the two BQA interviewing techniques are taken into account. In mathematical terms, by fitting the model on the data, we estimate the following conditional probability:

$$\begin{aligned}\Pr\left(Z_i = 1 \quad | \quad & \textbf{Age} \wedge \textbf{Education level} \wedge \textbf{Ethnicity} \wedge \\ & (\textbf{Employment status} \times \textbf{Employment status 2}) \wedge \\ & (\textbf{Mode} = \text{'BQA-CATI'} \vee \textbf{Mode} = \text{'BQA-M/W'})\end{aligned} \quad .$$

The reference point of reduced model (22) is defined by two demographic and two socioeconomic covariates. The categories of the demographic covariates are '$15-24$' (covariate **Age**) and 'First generation non-native' (covariate **Ethnicity**). The categories of the socioeconomic covariates are 'Low' (covariate **Education level**) and 'Act lf×Act lf' (covariate (**Employment status×Employment status 2**)).

The results of the effort that was undertaken to fit the reduced model on the measurement data, are presented in Table 25. It contains the set of model parameters accompanied by their estimated standard errors and the associated p-values. Inspection of the column with calculated p-values reveals that for two covariate categories, the change in value from their respective covariate reference

category is reckoned to be not significant. The two covariate categories in question are the category 'Missing' (covariate **Education level**) and the category 'Inact/non lf×Inact/non lf' (covariate (**Employment status**×**Employment status 2**)). For the remaining covariate categories, the change from reference value is deemed to be significant.

Next, in a discussion of the model fitting results, we focus our attention on the composite covariate (**Employment status**×**Employment status 2**). Generally speaking, a positive model parameter says, that relative to the conditional probability of the reference point, the conditional probability of encountering a responding panel respondent increases. A negative model parameter, in contrast, implies that the conditional probability of encountering a responding re-approached panel refuser increases with respect to the conditional probability of the reference point. For the two categories 'Act lf×Inact/non lf', 'Inact/non lf×Act lf', the (relative) increase is considerable by comparison.

In conclusion, the re-approach effort is generally successful in locating individuals who have undergone a change in labour force participation. More particularly, the re-approach effort is especially proficient when the individual in question moves out of the active labour force segment or when he/she moves into the active labour force segment out of a non labour force position. ■

Table 25. The fitted parameters of logistic regression model (22)

Covariates	*Parameter*	*St error*	*t-value*	*p-value*
Intercept	3.4411	0.4097	8.399	0.0000
Age				
$25 \leq$ age ≤ 54	−0.7967	0.3456	−2.297	0.0216
$55 \leq$ age ≤ 64	−1.3147	0.3458	−3.802	0.0001
Education level				
Average	0.3968	0.2104	1.886	0.0593
High	0.5854	0.2482	2.358	0.0184
Missing	0.2774	1.0570	0.262	0.7930
Ethnicity				
Second generation	0.8739	0.4743	1.843	0.0654
Natives	0.6958	0.2477	2.809	0.0050
Empl status×Empl status 2				
Act lf × Inact/non lf	−2.5223	0.3078	−8.196	0.0000
Inact/non lf × Act lf	−2.0361	0.3579	−5.689	0.0000
Inact/non lf × Inact/non lf	−0.2987	0.2146	−1.392	0.1639

5 Summary and conclusions

5.1 Summary

In 2005, a follow-up effort was undertaken for those households who refused to take part in the regular LFS panel of April to July. The interviewing techniques used in this follow-up or re-approach effort, are part of either the call-back approach (also known as the CBA-regime), or the basic question approach (also known as the BQA-regime). Within the CBA-regime we make distinction between interviews recorded via the CATI-technique or the CAPI-technique. Similarly, within the BQA-regime we allow interviewers to use the CATI-technique or web or mail questionnaires.

In this study, we have investigated the questions whether the CBA-regime and the BQA-regime were successful in getting interviewee's to respond or not, and whether LFS panel refusers act differently in the labour market than regular LFS respondents, or not.

5.2 Conclusions

From the results of the re-approach effort, we conclude that the type of interviewing technique used is largely responsible for obtaining a response from the interviewee. Apparently, the interviewee's labour market behavior is not significant in explaining whether or not he/she will respond. It follows then, that the response from the re-approach is a fair and accurate reflection of the labour market behavior of the LFS panel refusers.

When re-approaching LFS panel refusers, interviewers get a higher response rate when operating under the CBA-regime than when operating under the BQA-regime. Furthermore, the interviewer assisted interviewing techniques are the most effective interviewing techniques for getting a response, whereas a mail and web survey is the least effective interviewing technique for getting a response.

The re-approach effort appears to be successful in locating respondents whose position in the labour market has changed in the time since the last panel interview. This is true irrespective of the interviewing technique employed in the re-approach effort. The group of respondents reached by the re-approach effort is, therefore, a new and valuable extension to the group of regular panel respondents. It also means that the non-response effects in the LFS panel due to panel refusal, cannot be compensated for by means of weighting.

5.2.1 Conclusions regarding research question 1

The overall conclusion of the analysis conducted on behalf of research question 1 is, that the covariate **Mode** is the most significant covariate in predicting the

response behavior of an individual interviewed in the re-approach effort. Apparently, the used interviewing technique is chiefly responsible for successfully eliciting a response from an interviewee.

Response rates among the LFS panel refusers are quite high for households interviewed under the CBA-regime. Combined with the small number of covariates that relate to the re-approach response under this regime, we conclude that the CBA interviewing techniques give a good reflection of the LFS panel refusers 'employment status'. For further, more detailed conclusions, we address the case situations individually.

CBA+BQA There are strong levels of association between response variable R_i and the auxiliary variables **Mode** and **Age**. Furthermore, the remaining auxiliary variables under consideration have been found not associate with the response variable significantly.

A logistic regression modelling effort has demonstrated that for prediction of the response behavior of a re-approached individual, the covariate **Mode** is significant. None of the demographic or socioeconomic or geographic covariates has proven to be significant for the prediction of the response behavior.

CBA Strong levels of association exist between response variable R_i and the auxiliary variables **Ethnicity**, and **Mode**. Moreover, no significant level of association has been found to exist between the response variable and either of the remaining auxiliary variables.

Looking at the results of the logistic regression model building effort, we conclude that covariates **Ethnicity** and **Mode** are the only covariates considered to be significant for prediction of the response behavior of a re-approached individual. It should be taken into account, however, that the considered probability is a probability conditional on the fact that the interviewee is interviewed under the CBA-regime.

BQA Analysis has shown that strong levels of association exist between response variable R_i and the auxiliary variables **Mode**, **Ethnicity**, and **Age**. For the other auxiliary variables, the the association with R_i is not found to be significant.

The outcome of the logistic regressions analysis tells us that as far as the prediction of the response behavior is concerned, only covariate **Mode** is significant. The calculated p-values indicate that the contributions of the other covariates are not significant. When we consider the consequences of this conclusion, we should make allowances for the fact that the probability under consideration is a probability conditional on the fact that either of the two BQA interviewing techniques is used.

5.2.2 Conclusions regarding research question 2

The overall conclusion drawn on behalf of research question 2 states that the covariate (**Employment status**×**Employment status 2**) is the most significant covariate in predicting a respondent's origin in terms of he or she being a regular panel participant or a re-approached panel refuser. The actual figures show that the re-approach effort is especially useful in locating individuals who have experienced a recent change in the labour market.

This general conclusion indicates that the non-response due to panel refusal, follows a *Not-Missing-At-Random* (NMAR) pattern. This holds especially true for the situation in which we only take interviews under the CBA-regime and where the response rate is 76%. Hence, weighting will not be able to adjust for this effect. We address, again, the case situations individually.

CBA+BQA There exist strong levels of association between variable Z_i (the 0-1 indicator for response in LFS versus follow-up study) and the auxiliary variables **Employment status 2**, **Employment status**, **Education level**, **Ethnicity**, **Age** and **Household type**. It is only for the covariate **Urbanization** that we cannot reject the null-hypothesis of independence.

The results of the logistic regression analysis show that the covariates (**Employment status**× **Employment status 2**), **Ethnicity**, **Education level**, **Age**, and **Household type** are all significant for the prediction of the outcome of variable Z_i. Actual parameter values indicate that the regular LFS panel is loosing out on those individuals who have undergone a recent change in their position in the labour market.

CBA As in the Case 0 situation, strong levels of association exist between Z_i and the auxiliary variables **Employment status 2**, **Employment status**, **Education level**, **Ethnicity**, **Age** and **Household type**. In contrast to this, the null-hypothesis of independence cannot be rejected for the auxiliary variable **Urbanization**.

The logistic regression analysis of variable Z_i in this case situation, has results that are largely similar to those obtained in the Case I situation. This implies that, once more, we find the covariates (**Employment status**×**Employment status 2**), **Ethnicity**, **Education level**, and **Household type** to be the covariates that are all significant for the prediction of the behavior of the variable.

BQA Strong levels of association exist between variable Z_i and the auxiliary variables **Employment status 2**, **Employment status**, **Age**, **Ethnicity** and **Education level**. For the two variables **Urbanization** and **Household type**, the hull-hypothesis of independence cannot be rejected.

It was revealed in the logistic regression analysis of target Z_i that in this case situation, the behavior of the variable is best explained by the significant covariates (**Employment status**×**Employment status 2**), **Age**, **Ethnicity**, and **Education level**. According to the model data, the re-approach effort is proficient in finding people who are moving in and / or out the active labour force market segment.

References

Agresti, A. (2002), *Categorical Data Analysis; second edition*, ISBN 0-471-36093-7, 2002.

Banning, R. (2008), *Data preparations for a study on the re-approach of LFS panel refusers*, in preparation, Statistics Netherlands, 2008.

Cobben, F., and B. Schouten (2007a), *A follow-up with basic questions of non-respondents to the Dutch LFS*, DP 07011, Statistics Netherlands, 2007.

Cobben, F., and B. Schouten (2007b), *Are you the next to have your birthday? congratulations: you may answer some questions!*, DMV-2007-01-09-FCBN, Statististics Netherlands, 2007.

Kersten, H.M.P., and J.G. Bethlehem (1984), *Exploring and reducing the non-response bias by asking the basic question*, BPA 627-84-M1, CBS, Voorburg, Statistics Netherlands, 1984.

Koch, G.G., D.H. Freeman, and J.C. Freeman (1975), *Strategies in the multivariate analysis of data from complex surveys*, International Statistical Review, 43, pp 59-78, 1975.

Rao, J.N.N., and A.J. Scott (1984), *On Chi-squared Tests for Multiway Contingency Tables with Propertions Estimated From Survey Data* Analys of Statistics, 12, pp 46-60, 1984.

Schouten, J.G. (2007), *A follow-up of nonresponse in the Dutch Labour Force Survey*, Statistics Netherlands, 2007.

Appendix A Definitions of auxiliary variables

Variable **Household type**: Single (person); Single + (single parent with minors or single parent with minors and others or single person with others); Couple; Couple + (couple with minors or couple with minors and others or couple with others).

Variable **Ethnicity**: First generation non-native; Second generation non-native; Native.

Variable **Age**: 15 $\leq$age$\leq$ 24; 25 $\leq$age$\leq$ 54; 55 $\leq$age$\leq$ 64.

Variable **Employment status**: Active labour force; Inactive labour force; Non labour force.

Variable **Education level**: Low; Average; High.

Variable **Urbanization**: Very strong; Strong; Moderate; Little; Not.

Appendix B Frequency distributions of the auxiliary variables

Table 26. Frequency distributions of the demographic variables for the different case situations under research question 1

	Case 0		Case I		Case II	
Variable	*Freq*	*Perc*	*Freq*	*Perc*	*Freq*	*Perc*
Household type						
Single	216	28.3	108	28.4	108	28.2
Single +	81	10.6	37	9.7	44	11.5
Couple	228	29.9	115	30.3	113	29.5
Couple +	238	31.2	120	31.6	118	30.8
Total	763	100.0	380	100.0	383	100.0
Ethnicity						
First generation	152	19.9	68	17.9	84	21.9
Second generation	53	7.0	28	7.4	25	6.5
Natives	558	73.1	284	74.7	274	71.6
Total	763	100.0	380	100.0	383	100.0
Age						
$15 \leq$ age ≤ 24	98	12.8	56	14.7	42	11.0
$25 \leq$ age ≤ 54	430	56.4	200	52.6	230	60.1
$55 \leq$ age ≤ 64	235	30.8	124	32.6	111	29.0
Total	763	100.0	380	100.0	383	100.0

Table 27. Frequency distributions of the socioeconomic variables for the different case situations under research question 1

	Case 0		Case I		Case II	
Variable	*Freq*	*Perc*	*Freq*	*Perc*	*Freq*	*Perc*
Employment status						
Active labour force	374	49.0	176	46.3	198	51.7
Inactive labour force	30	3.9	20	5.3	10	2.6
Non labour force	359	47.1	184	48.4	175	45.7
Total	763	100.0	380	100.0	383	100.0
Education level						
Low	339	44.4	174	45.8	165	43.1
Average	265	34.8	132	34.7	133	34.7
High	149	19.5	71	18.7	78	20.4
Missing	10	1.3	3	0.8	7	1.8
Total	763	100.0	380	100.0	383	100.0

Table 28. Frequency distribution of the geographic variable for the different case situations under research question 1

	Case 0		Case I		Case II	
Variable	*Freq*	*Perc*	*Freq*	*Perc*	*Freq*	*Perc*
Urbanization						
Very strong	159	20.8	78	20.5	81	21.1
Strong	219	28.7	105	27.6	114	29.8
Moderate	153	20.1	80	21.1	73	19.1
Little	142	18.6	72	19.0	70	18.3
Not	90	11.8	45	11.8	45	11.7
Total	763	100.0	380	100.0	383	100.0

Table 29. Frequency distributions of the demographic variables for the different case situations under researach question 2

	Case 0		Case I		Case II	
Variable	*Freq*	*Perc*	*Freq*	*Perc*	*Freq*	*Perc*
Household type						
Single	1,959	11.3	1,919	11.2	1,876	11.0
Single +	1,279	7.4	1,266	7.4	1,248	7.3
Couple	4,744	27.4	4,684	27.3	4,655	27.3
Couple +	9,346	53.9	9,285	54.1	9,258	54.3
Total	17,328	100.0	17,154	100.0	17,037	100.0
Ethnicity						
First generation	1,356	7.8	1,328	7.7	1,300	7.6
Second generation	1,012	5.8	1,005	5.9	985	5.8
Natives	14,960	86.3	14,821	86.4	14,752	86.6
Total	17,328	100.0	17,154	100.0	17,037	100.0
Age						
$15 \leq$ age ≤ 24	2,876	16.6	2,863	16.7	2,830	16.6
$25 \leq$ age ≤ 54	11,079	63.9	10,984	64.0	10,934	64.2
$55 \leq$ age ≤ 64	3,373	19.5	3,307	19.3	3,273	19.2
Total	17,328	100.0	17,154	100.0	17,037	100.0

Table 30. Frequency distributions of the socioeconomic variables for the different case situations under research question 2

	Case 0		Case I		Case II	
Variable	*Freq*	*Perc*	*Freq*	*Perc*	*Freq*	*Perc*
Employment status						
Active labour force	11,436	66.0	11,347	66.1	11,303	66.3
Inactive labour force	667	3.8	664	3.9	653	3.8
Non labour force	5,225	30.2	5,143	30.0	5,081	29.8
Total	17,328	100.0	17,154	100.0	17,037	100.0
Employment status 2						
Active labour force	11,467	66.2	11,387	66.4	11,334	66.5
Inactive labour force	549	3.2	541	3.2	536	3.1
Non labour force	5,312	30.7	5,226	30.5	5,167	30.3
Total	17,328	100.0	17,154	100.0	17,037	100.0
Education level						
Low	5,721	33.0	5,646	32.9	5,580	32.8
Average	7,004	40.4	6,943	40.5	6,910	40.6
High	4,531	26.1	4,494	26.2	4,478	26.3
Missing	72	0.4	71	0.4	69	0.4
Total	17,328	100.0	17,154	100.0	17,037	100.0

Table 31. Frequency distribution of the geographic variable for the different case situations under research question 2

	Case 0		Case I		Case II	
Variable	*Freq*	*Perc*	*Freq*	*Perc*	*Freq*	*Perc*
Urbanization						
Very strong	2,959	17.1	2,929	17.1	2,865	17.0
Strong	4,744	27.4	4,695	27.4	4,663	27.4
Moderate	4,043	23.3	4,002	23.3	3,980	23.4
Little	3,386	19.5	3,351	19.5	3,335	19.6
Not	2,196	12.7	2,177	12.7	2,164	12.7
Total	17,328	100.0	17,154	100.0	17,037	100.0

Table 32. Frequency distributions of the composite covariate for the different case situations under research question 2

Variable	**Employment status 2**			
	Active lab force	Inactive lab force	Non lab force	Total
Case 0				
Employment status				
Act lf	11,056	83	297	11,436
Inactive lab force	140	296	231	667
Non lab force	271	170	4,784	5,225
Total	11,467	549	5,312	17,328
Case I				
Employment status				
Active lab force	10,989	78	280	11,347
Inactive lab force	140	296	228	664
Non lab force	258	167	4,718	5,143
Total	11,387	541	5,226	17,154
Case II				
Employment status				
Active lab force	10,948	81	274	11,303
Inactive lab force	135	289	229	653
Non lab force	251	166	4,664	5,081
Total	11,334	536	5,167	17,037

Table 33. Frequency distributions of the new collapsed composite covariate under research question 2

Variable	**Employment status 2**		
	Active labour force	Inactive/non labour force	Total
Employment status			
Active labour force	10,948	355	11,303
Inactive/non labour force	386	5,348	5,734
Total	11,334	5,703	17,037