Representativity of the LISS panel

09

Jan van der Laan

Discussion paper (09041)

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

6008309041 X-10

Representativity of the LISS panel

JAN VAN DER LAAN

Abstract: The goal of the MESS-project is to set up a longitudinal internet panel of households, the LISS panel. Academic researchers can submit proposals for specific research they want to perform with this panel. In order to obtain a representative panel, a traditional probability sample was drawn from a population register, and households without internet access or computer were offered free internet access and a simple to use computer. The analysis in this paper shows that certain groups are underrepresented in the LISS panel. Especially single person households, households with a high average age and households with first generation immigrants. We also show that offering a free computer and internet access has a positive effect on the representativity. Especially households with a high average age would otherwise be even more underrepresented. We also compared the response of the LISS panel to that of the Labour Force Survey (LFS). This analysis showed that there are substantial differences in representativity. These are largely caused by differences in response for the two different modes (telephone and face-to-face interview) used in the recruitment of the panel. When the effect of mode is removed from the model, the representativity is much closer to that of the LFS.

Keywords: LISS, MESS, representativity, internet panel

Contents

1 Introduction

The LISS panel is a longitudinal internet panel of households, that is set up by the MESS-project (An Advanced Multi-Disciplinary Facility for Measurement and Experimentation in the Social Sciences). Statistics Netherlands is one of the participants in this project. For Statistics Netherlands it is of interest to see if it is possible to obtain accurate results using an internet panel since this could be a cost efficient method of data collection.

Although internet panels have the advantage of being relatively cheap, they are often not representative for the population (e.g. see Bethlehem, 2006). Most internet surveys use self selection, which means that anyone willing can join the panel. The researcher has little influence on who joins and who do not. Therefore, the respondents will consist mainly of internet users that are for some reason motivated to participate. In order to avoid the problems of self selection, the LISS panel uses a traditional probability sample drawn from population registers. This sample was provided by Statistics Netherlands.

Even if an internet panel were to use a traditional drawn sample, only households having access to the internet can participate. As not all groups in the target population have access to internet, this will cause a selective underrepresentation of certain groups. For example, in 2008 91% (Statistics Netherlands, 2008*a*) of the total Dutch population had internet access, while only 57% of the population aged 65 years or older had internet access. Internet surveys will therefore have an underrepresentation of elderly people. In order to avoid this problem in the LISS panel, households without internet access are offered free internet access and if necessary a simple to use computer is installed for them. In the MESS project special focus is given to elderly and immigrants since these are usually underrepresented in surveys and especially internet surveys.

In this paper we will investigate the representativity of the LISS panel. We will first describe the methods used to investigate the representativity. After looking at the representativity, the effect of offering internet access to households without internet access on the representativity will be investigated. Finally, the response of the LISS panel is compared to that of the Dutch Labour Force Survey.

2 Methods

2.1 Data collection

As was mentioned earlier, a traditional probability sample was used using registers as population frame. As a population frame the population registers (in Dutch: Gemeentelijke Basis Administratie) were used. Since the LISS panel is a household panel, a random sample of households was needed. However, since the population registers do not register households but only persons, a two step approach was used. First, addresses were selected randomly from the registers. For each address one person was then selected. During the recruitment of the households these persons were used as an entry into the household. The total sample contains 10,150 households.

The names and addresses of the selected persons were obtained from the population registers. To these were added any registered telephone numbers. Households with a registered telephone were contacted using telephone (CATI). The remainder and those who could not be contacted using telephone were visited by the interviewers (CAPI).

During the recruitment, sampled persons were first asked to participate in a interview, the recruitment interview. In this interview general background questions were asked about the respondent. At the end of the interview the respondent was told about the panel and asked if he or she would like to participate. Households without access to the internet, or who were worried that an internet survey might be too complicated for them, were told about the simple to operate computer with internet access that could be installed in their homes for free for the duration of the panel. To demonstrate the use of this computer they were shown a demonstration video.

People who did not want to participate in the recruitment interview were asked if they would at least answer a couple of key questions. They were also told about the panel and asked to participate in the panel. Finally, a small percentage of people who said they were willing to join the panel did not join in the end. Therefore, the response process can be split into a number of steps:

Contact Those who could be contacted for the recruitment interview.
Primary response Those who participated in the recruitment interview. This includes persons who were only willing to answer the three basic questions.
Secondary response Those who agreed to participate in the panel
Tertiary response Those who actually did participate in the panel

In every step a certain percentage of persons is lost. In order to investigate the representativity of the panel, only the tertiary response needs to be investigated. However, by investigating each of the steps a better understanding is obtained of the mechanisms that led to the observed response.

2.2 Data used to investigate representativity

In order to investigate the representativity, we first need to establish what we consider a representative sample. Unfortunately, there is not one single definition of representativity, see for example Schouten and Cobben (2007) for a discussion on representativity measures. The most natural choice would be that a survey is representative for a certain set of variables, when estimators based on these variables are unbiased. However, this would mean that the representativity needs to be investigated for every possible set of variables. Since the panel is going to be used for many different kinds of research this is not possible. Furthermore, in order to investigate the representativity for a certain set of variables, the values of these variables need to be known for both the respondents and non-respondents, which is clearly not possible in practice. Therefore, we will investigate the representativity for a set of variables known from registers for both respondents and non-respondents.

As we will investigate the response probability as a function of certain background variables known for both respondents and non-respondents, we will need to use variables from registers. Since the LISS panel is a household panel, we need background

properties of the households. For that purpose, we used the data from the Household Statistic of Statistics Netherlands. See Harmsen and Israëls (2003) for a description of the methods used to create the Household Statistics. This data set contains demographic information for every household member of every household in the Netherlands. This information is obtained from the municipal registers[1].

As was just mentioned, the representativity is investigated using background properties obtained from registers for both responding and non-responding households. Therefore, the sampled households were first coupled to the household data using the address information available in the LISS data files (postal code, house number, additional house number). When only one household is living at the address this uniquely identifies the household. For 2.3% of the households in the sample, the address could not be matched to an address in the household data. This is probably caused by errors in the address. Especially the additional house number is not always coded consistently in both files, e.g. one file uses 'IV' while the other file uses '4'[2]. When more than one household lives at the address, the date of birth of the person participating in the interview (only available for those households that participated in the recruitment interview) was used to further identify the correct household. For 97.8% of the coupled households in the sample, the corresponding household in the household data can be uniquely identified. For the remaining households, all persons living at the address are considered to be part of the household. The type of household for these households was set to 'unknown'.

Using the coupled data from the household data file, aggregate variables were generated for each household in the sample. Table 2 shows the variables that were generated. 'Type of household' is a categorical variables with the following categories: 'Single person household', 'Unmarried couple without children', 'Married couple without children', 'Unmarried couple with children', 'Married couple with children', 'Single parent household', 'Other', and 'Unknown'.

The ethnicity of the household is used to distinguish between the different ethnic backgrounds of immigrants. Someone is considered an immigrant when he or she or one of the parents is born outside of the Netherlands. A first generation immigrant is born outside of the Netherlands; a second generation immigrant has one of his or her parents born outside of the Netherlands. Table 1 shows the definition of ethnicity for the households. The households are further divided by the generations living in the households: only first generation, first and second generation, and only second generation.

Besides demographic background variables of the households, we also used background information of the neighbourhood in which the household lives. The neighbourhood can be used as a predictor of the economic and social status of the household. These variables were obtained from StatLine (Statistics Netherlands, 2005/2007). Ta-

[1] Using data from the municipal registers the occupants are obtained for every address. These are then grouped into households using the relations that exist between the persons. For example, father, mother and children are combined into one household. For roughly 93% of the addresses the households can be determined uniquely. For the remaining 7% of the addresses the households are imputed using background variables of the persons. For example, a man and woman approximately the same age living at one address are imputed as a pair with a high probability.

[2] In the case of Roman numerals versus Arabic numerals and uppercase and lowercase letters, both options were allowed during coupling.

Table 1: Definition of household ethnicity. Households are first divided into non-western, western and non-immigrant; non-western is further divided into subcategories. Ethnicity is defined by the first rule that applies.

Ethnicity	Condition
Non-western	At least one non-western immigrant in the household
Morocan	At least one Morocan immigrant in the household
Turkish	At least one Turkish immigrant in the household
Surinamese	At least one Surinamese immigrant in the household
Netherlands Antilles	At least one immigrant from the Netherlands Antilles
Other	Otherwise
Western	At least one western immigrant in the household
Dutch	Otherwise

Table 2: Background variables for the households

Household properties
Type of household
Number of household members
Number of households on the address
Ethnicity of household members
Average age of adults aged 18 years or older
Gender of household member[a]
Number of minors[b]
Age youngest child
Neighbourhood properties
Urbanisation
Percentage of non-western immigrants
Percentage of western immigrants
Average property value
Average number of cars per household
Average number of household members
Percentage of single person households
Percentage of rented houses[c]
Percentage of people receiving social benefits[c]
Percentage of people with a high income[c]
Percentage of people with a low income[c]
Percentage of people employed[c]

[a] Only available for single person households

[b] Only available for 'married/unmarried couple with children' and 'Single parent household'

[c] Based on 2005 data

ble 2 shows the background variables of the neighbourhoods. For some variables the most recent version was 2005 instead of 2007. For 5.3% of the households the neighbourhood properties from 2005 are missing, since the code of the neighbourhood in 2007 could not be coupled to the neighbourhoods of 2005.

2.3 Univariate analysis

The representativity was first analysed using univariate methods. Each of the variables in table 2 was recoded into categorical variables as far as they were not already. The households are divided into groups using these categories and the response rates (number of households responding divided by the total number of households in the sample in the category) in the groups are compared to each other. In case of a representative response, the response rates should be equal.

The categories for a given continuous variable are defined using 10% quantiles, which gives 10 categories for each variable. The first category contains the 10% of households with the smallest values, and the tenth category contains the 10% of households with the largest values of the variable.

The significance of the differences was investigated using a contingency table were the categorical variable is crossed with the tertiary response. A χ^2-test using a p-value of 0.05 was used. When an effect is significant it is also interesting to see the magnitude and direction of the effect. Therefore, the contingency tables are displayed using spine plots (e.g. see figure 1). The response in each category of a property is shown in a vertical bar. The width of a bar corresponds to the number of households in the category. The bars are then further divided according to the type of response: 'NC' (non-contact), 'NR1' (refusal to participate in the recruitment interview), 'NR2' (refusal to participate in the panel'), 'NR3' (no participation in the panel after agreeing to participate in the panel), and 'R' (tertiary response).

2.4 Multivariate analysis

As many of the variables in table 2 are correlated with each other, we also performed multivariate analysis. For example, age and type of household are strongly correlated. It might be possible that the differences in response seen for the different types of household are completely explained by the differences in the age distribution of the different types of household. In this case we will end up with a model that only includes one of these parameters: the one that predicts the response best.

We used a logistic model using the variables in table 2 as predictors for the response. The better this model predicts the observed response, the larger the dependence of the response on the covariates, and the larger the non-representativity with respect to these variables.

Besides a model for the total tertiary response, we also estimated conditional models for each of the steps involved in the response: contact, primary response given contact, secondary response given primary response, and tertiary response given secondary response. These can be used to gain insight into the mechanisms involved in the non-response.

In order to deal with the fact that some neighbourhood properties are not available for all households, a stratified model was estimated for the neighbourhood properties. For those households for which the neighbourhood properties are available a model was estimated with these properties. For those households where the properties are not available we estimated only an intercept. Let δ_{05} be an indicator that is one if all neighbourhood properties from 2005 are available for household i and zero otherwise. δ_{07} is defined likewise. The model that is estimated can be written as

$$\text{logit}(p_i) = \alpha + X_{hh}\beta_{hh} + \alpha_{05}\delta_{05} + \alpha_{05}X_{05}\beta_{05} + \alpha_{07}\delta_{07} + \alpha_{07}X_{07}\beta_{07}, \qquad (1)$$

where α is the intercept, X_{hh}, X_{07} and X_{05} are the vectors containing the observed household and neighbourhood properties, β_{hh}, β_{07} and β_{05} are the corresponding parameter vectors.

The variables 'Number of household members', 'Type of household', and 'Number of minors' are dependent on each other. In order to facilitate interpretation of the model a recoded variable 'Number of additional adults' was created, which is defined as the number of adults minus the minimum number of adults for the given type of household. The minimum number of adult members for most types of households is two, except for 'single parent' and 'single person' where the minimum number is one. Therefore, the regression parameters of for example 'married couple with children' corresponds to a household consisting of two adults. In most cases the 'additional adults' are adult children living at home.

3 Representativity of the LISS panel

3.1 Response rates

Table 3 shows the response in each of the four steps. In the end 46% of the sampled households participate in the panel. In each of the steps roughly 10 to 20% of the households remaining from the previous step are lost. Especially in the third step, secondary response, a large fraction of households is lost. Approximately 25% of the households that are willing to participate in the recruitment interview are not willing to participate in the panel. This could be because they find the burden of participating in a panel too large, or because they do not have internet access or a computer and cannot be persuaded by offering the simple to use computer. The percentage of non-contact (approx. 14%) is quite high.

Table 3: Response rates. The first column shows the fraction of households in the sample participating in each of the steps. The second column shows the response conditional on participation in the previous step.

	Total	Conditional
Contact	85.7%	85.7%
Primary response	69.3%	80.9%
Secondary response	51.4%	74.1%
Tertiary response	45.9%	89.4%

It should be noted that the response fractions shown here somewhat differ from the official numbers published by CenterData since we used the raw response data. In the official numbers, the approximately 4% of the non-response originating from incorrect addresses etc. is removed from the data set resulting in a non-contact of 10% and a tertiary response of 48%.

3.2 Univariate analysis

Response rates by themselves are not an effective way to measure representativity. Since the LISS panel is an internet panel, there might be a significant underrepresentation of certain groups of households. In this section the sampled households are divided in groups using different background properties. We then compare the response rates of the groups to each other. Table 2 shows the properties that were investigated. All of the background properties discussed below show a significant effect (see paragraph 2.3).

Figure 1 shows the response as a function of the average age of the adult household members aged 18 years or older. At first, the response increases with age. It then reaches a maximum between 40 and 50 years and decreases after that. The reason for the nonresponse clearly differs for younger households and older households. For younger households the nonresponse is mainly caused by non-contact. The nonresponse for older households is caused mainly by refusal to participate in the panel. This is probably caused by the fact that these households do not have a computer with internet access. The fact that they are offered a simple to use computer with internet access, does not completely solve this. They might be put off by the fact that they have to learn how to use a new technology, or they are unwilling to commit themselves for a longer time to a panel.

There are large differences in response for the different types of households, as is shown in figure 2. The response for single person households is much lower than that of couples. This could be caused by the fact that the probability of finding at least one household member at home increases when the number of household members increases. An other reason could be the relatively high age of these households (54 years), which also explains why female single person households have a lower response than male (not shown in the figure).

The response is highest for couples with children. The fact that they are married or not has little influence. The average age of the adults in this group is in the range 30–50 years, which correlates with the higher response rates in these age categories. The unmarried couples without children have a relatively high non-contact rate. Since a large number of these couples are aged below 30, this correlates with the higher non-contact rate for younger households. For unmarried couples it will probably also more freqently occur that both members work, making contact more difficult. The married couples without children are relatively old (on average 60 years). These are probably for a large part couples whose children have left the house.

The response for the different ethnic background of immigrants is shown in figure 3. The response for the different non-western immigrant groups is lower than that of the households with exclusively native persons except for households with members

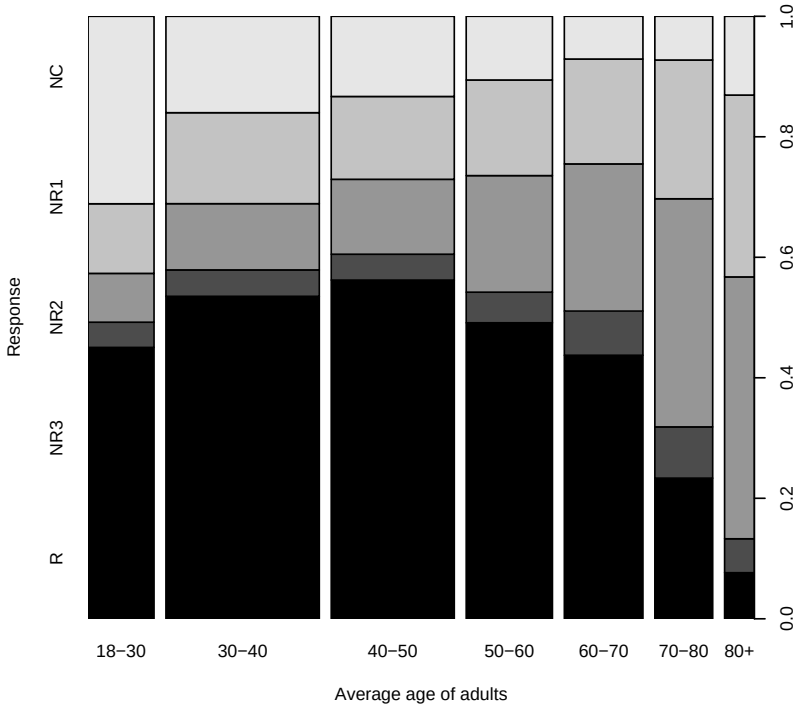


Figure 1: Response as a function of the average age of the household members of 18 years and older.

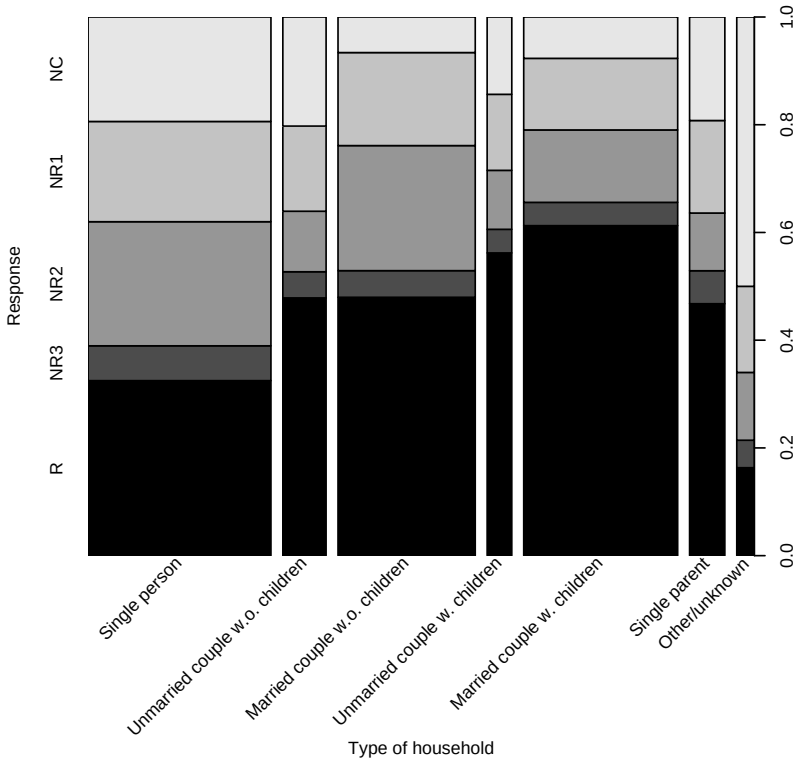


Figure 2: Response for the different types of households

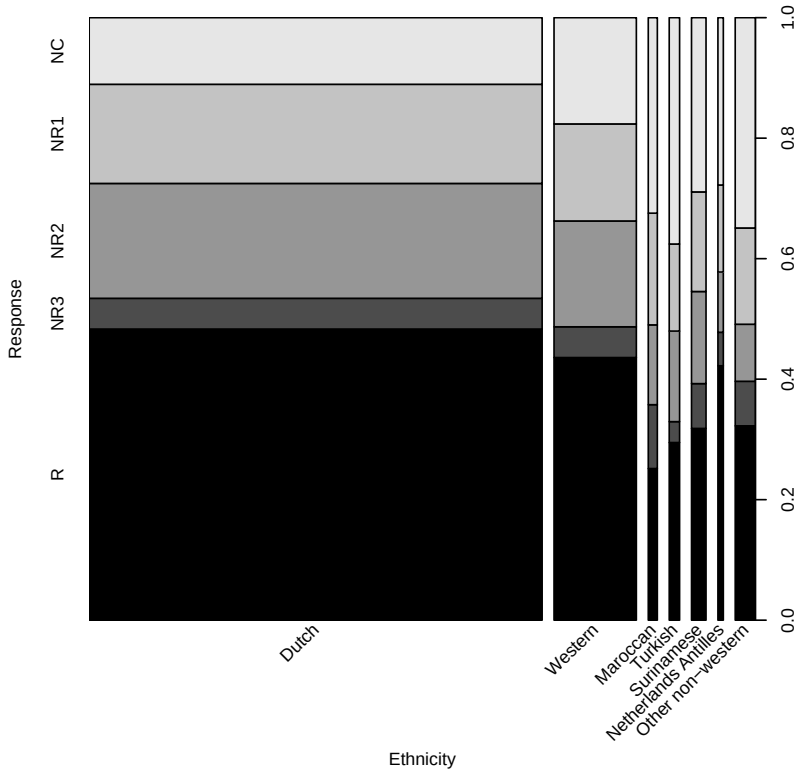


Figure 3: Response for the different ethnic groups. A household belongs to one of the non-native groups when at least one person in the household is an immigrant.

from the Netherlands Antilles. The fact that this last group has the same response as the native households, suggests that language problems might be a cause for the lower response for the other groups. This is backed up by the fact that the response rate for households with exclusively second generation immigrants is equal to that of households with native Dutch members as is shown in figure 4. However, only in 12% of the cases the reason for non-contact in case of immigrant households was marked as 'language problems'. So either language problems are not always marked, or there are also other factors that play a role in the high non-contact rate for immigrants.

The effect on the response of the background properties of the neighbourhood in which the household lives, is shown is figures 5 to 9. As can be seen in figure 5, the response is lower for high and very high urbanisation rates. The other figures all show the same effect: the response decreases as the indicators of the social status of the neighbourhood decrease.

From the univariate analyses it can be concluded that response is selective: response is lower for single person households, old and young households, immigrants, households living in highly urbanized neighbourhoods, etc.

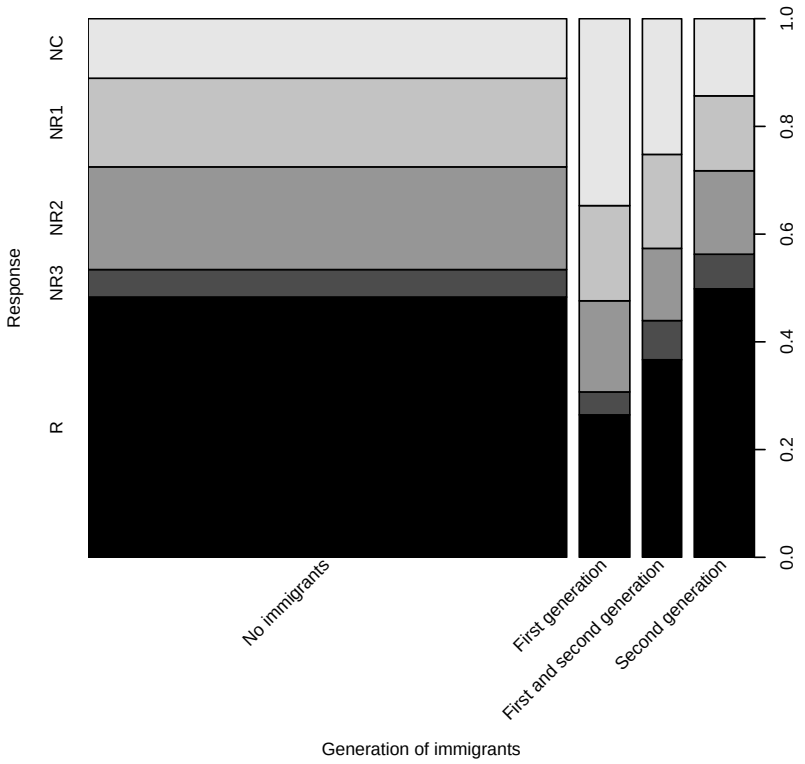


Figure 4: Response for the different generations of immigrants in the household.

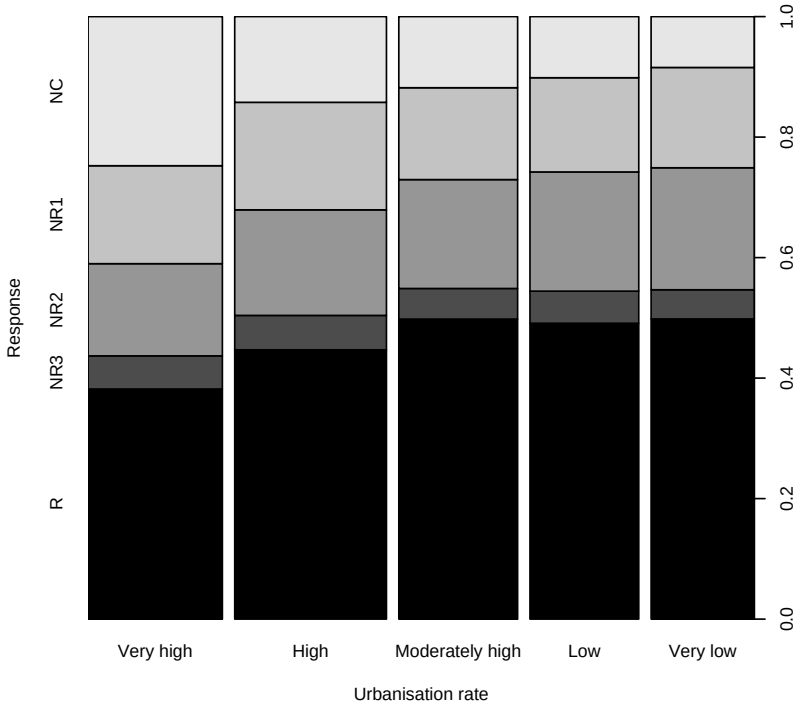


Figure 5: Response as a function of urbanisation.

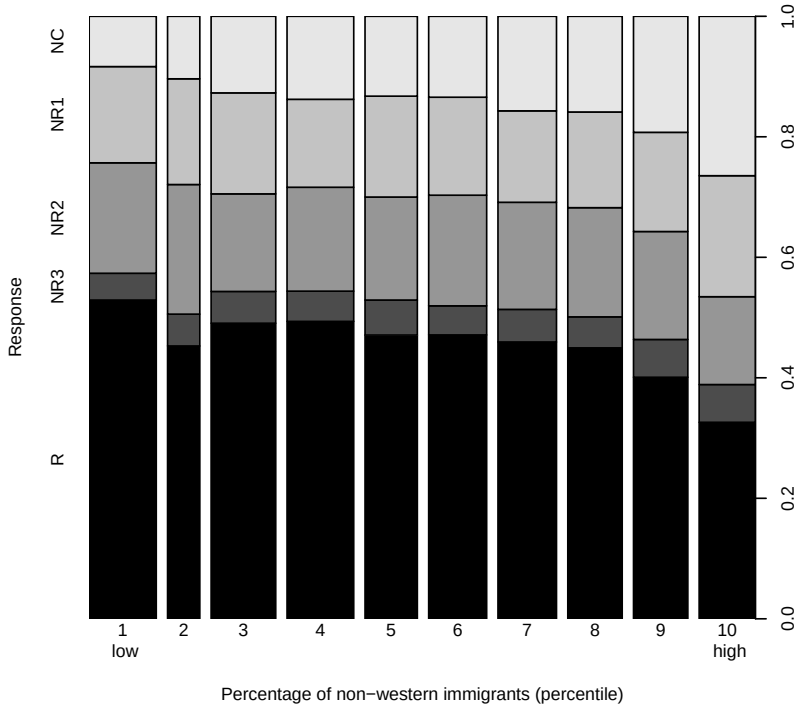


Figure 6: Response for different categories of the percentage of non-western immigrants in the neighbourhood.

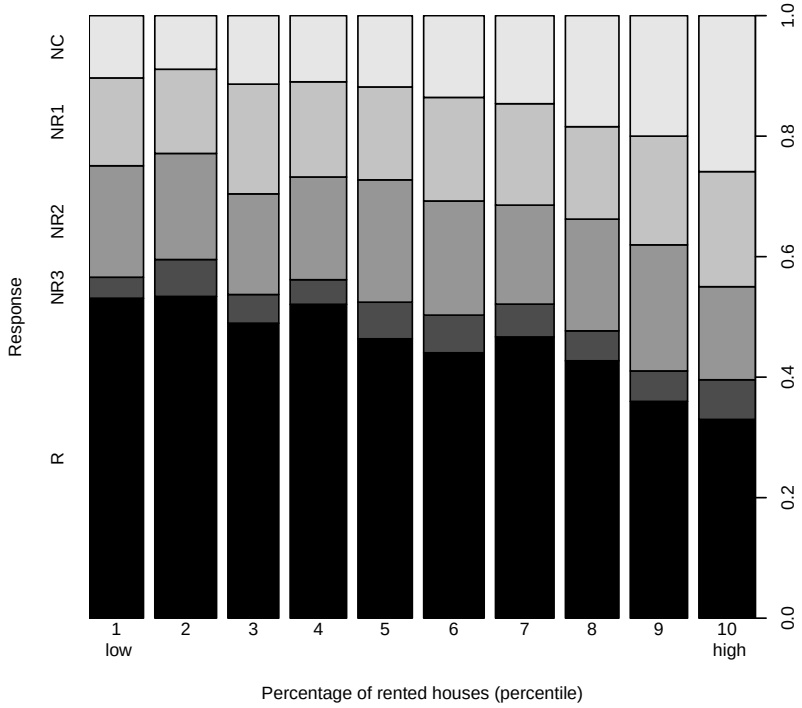


Figure 7: Response for different categories of the percentage of rented houses in the neighbourhood.

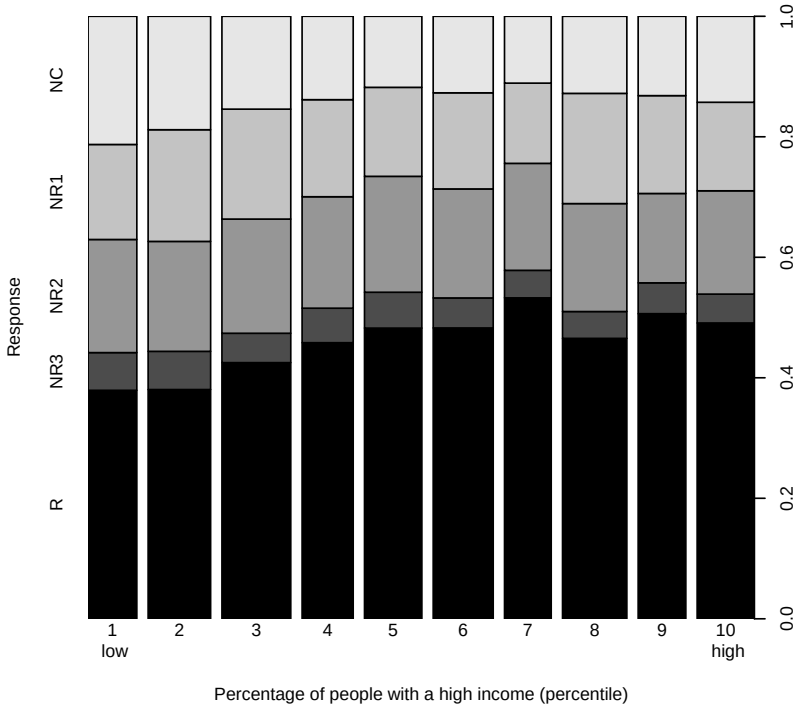


Figure 8: Response for different categories of the percentage of people with a high income in the neighbourhood.

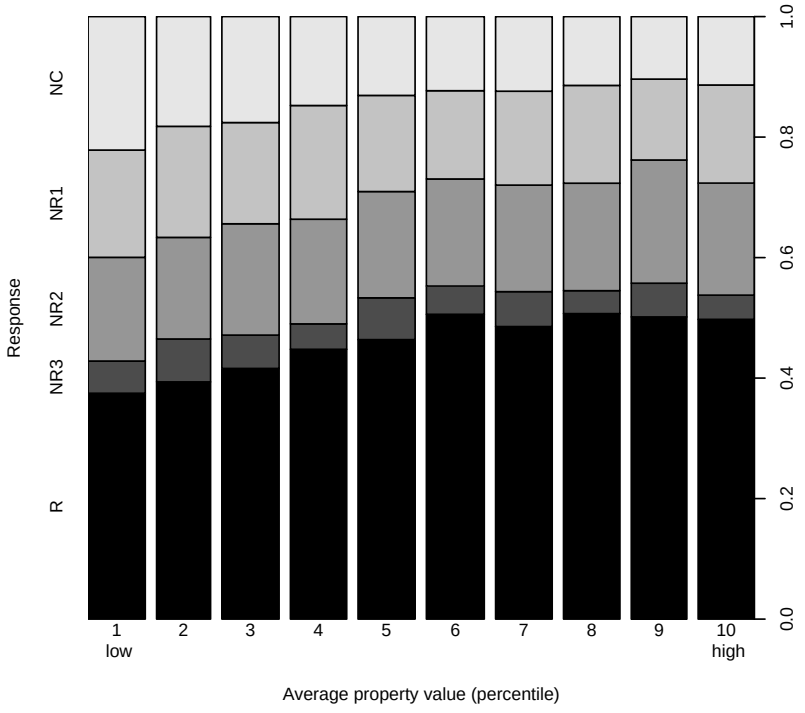


Figure 9: Response for different categories of the property value in the neighbourhood.

3.3 Multivariate analysis

3.3.1 Overall response

In order to investigate the representativity, a model is estimated that tries to predict the response using the available background properties. As was discussed in section 2.4 a logistic model is used for this. The better this model is able to predict the response, the stronger the response depends on the background properties and the worse the representativity is. For example, a positive regression coefficient for age means that the response increases with age and, therefore, an overrepresentation of older households.

In the previous section, is was shown that the response first increases with age (the average age of the adult household members) and then starts to decrease again with increasing age (see figure 1). Therefore, the effect of age could not be modelled using a simple linear model. We used a second degree polynomial. The fit of the model using a polynomial was equal to that of a model using a categorical variable for age but uses less parameters. To aid in interpreting the two parameters for age, $\beta_{\text{(Average age)}}$ and $\beta_{\text{(Average age squared)}}$, the table also shows the average age at which the response reaches a maximum. This maximum follows from the two parameters (the x-coordinate of the maximum of the quadratic function $a+bx+cx^2$ is given by $x_{max} = -b/2c$):

$$\text{(Maximum for average age)} = \frac{-\beta_{\text{(Average age)}}}{2\,\beta_{\text{(Average age squared)}}}.$$

In order to estimate this model, we started with a model including all parameters including the following interaction terms: Type of household $\times$ Ethnicity and Single person household $\times$ Gender. These were added, because we suspected that the effect of, for example, ethnicity on response might differ for the different types of household. One by one parameters that were not significant at the 95% level were removed from the model (a likelihood ratio test was used for this). At each step we also tested if any significant parameters could be added to the model using a likelihood ratio test.

Table 4 shows the estimated parameters for the logistic model for the tertiary response. Only the significant parameters are included in the table. The response of all types of household except 'other/unknown' is higher than that of the reference category of single person households, which was also visible in the univariate analysis. When we adjust for all other covariates, we see that the response for married couples is higher than that of unmarried couples. This result differs somewhat from that of the univariate analysis, where the response for married and unmarried couples was approximately the same. However, as was discussed there, there are large differences in age between the married and unmarried couples. The response decreases with the number of additional adults (aged 18 years or older) in the household. These are often adult children who are perhaps less willing to answer questions about the household they are part of.

The results of the multivariate analysis confirm that the households with solely second generation immigrants have about the same response as households with solely native Dutch members. The ethnic background of the immigrants was not significant in the model. The differences between the different ethnic backgrounds seen in the univariate analysis (see figure 3), are therefore probably caused by differences in for example

Table 4: Estimated parameters for logistic model for tertiary response. Values with a p-value below 0.05 are marked with ''.*

Parameter	Estimate
Intercept	-2.21*
Household properties	
Household type=Single person	0[a]
Household type=Unmar.couple w/o children	0.198*
Household type=Mar.couple w/o children	0.569*
Household type=Unmar.couple w.children	0.33*
Household type=Mar.couple w.children	0.515*
Household type=Single parent	0.31*
Household type=Other/unknown	-1.28*
Ethn.=Native Dutch	0[a]
Ethn.=First generation western	-0.48*
Ethn.=First+second generation western	-0.5*
Ethn.=Second generation western	-0.018
Ethn.=First generation non-western	-0.83*
Ethn.=First+second generation non-western	-0.67*
Ethn.=Second generation non-western	0.29
Number of households at address	0.306*
Average age	0.0704*
Average age squared	-0.001094*
Registered telephone	1.678*
Number of additional adults	-0.282*
Neighbourhood properties	
Non-missing 2007	0.45*
Average property value neighbourhood	-0.00089*
Non-missing 2005	0.53*
Percentage rented houses neighbourhood	-0.0042*
Percentage w.low income neighbourhood	-0.01*
Maximum for average age	32.2

[a] Reference category

generation and age.

When corrected for other variables such as type of household, the response reaches a maximum for an average age of the household members of approximately 32 years after which it starts to decline.

3.3.2 Conditional response

In order to get more insight into the processes involved in the non-response, we also estimated conditional models for each of the steps involved. For example, a model was estimated that predicts the probability of primary response given contact. We used the same model as for the total tertiary response from the previous section, but added a variable the codes the ethnic origin of immigrants. This because participation of immigrants is an important goal of the MESS project. These models are summarized in table 5.

Table 5: Parameter estimates for the conditional models for each of the steps in the response process. Values with a p-value below 0.05 are marked with ''.*

Parameter	Contact	Primary	Secondary	Tertiary
Intercept	-0.78*	1.35*	0.72	1.58*
Household properties				
Household type=Single person	0[a]	0[a]	0[a]	0[a]
Household type=Unmar.couple w/o children	0.41*	-0.14	0.19	0.16
Household type=Mar.couple w/o children	0.88*	0.02	0.213*	0.86*
Household type=Unmar.couple w.children	0.71*	-0.05	0.09	0.36
Household type=Mar.couple w.children	1.29*	0.00	0.00	0.47*
Household type=Single parent	0.53*	-0.16	0.41*	0.09
Household type=Other/unknown	-0.75*	-0.74*	-0.96*	-0.55
Ethn.=Native Dutch	0[a]	0[a]	0[a]	0[a]
Ethn.=First generation western	-0.95*	-0.16	-0.33	0.51
Ethn.=First+second generation western	-0.82*	-0.34	-0.22	-0.27
Ethn.=Second generation western	-0.14	0.13	0.08	-0.31
Ethn.=First generation non-western	-1.12*	-0.21	-0.24	-0.88*
Ethn.=First+second generation non-western	-1.09*	-0.07	-0.04	-1.13*
Ethn.=Second generation non-western	-0.3	0.53	0.98*	-0.21
Number of households at address	0.132*	0.149*	0.48*	0.084
Average age	0.07*	-0.008	0.018	0.046*
Average age squared	-0.00045*	-0.00013	-0.00080*	-0.00089*
Fixed telephone	0.39*	2.017*	2.019*	0.37*
Number of additional adults	-0.067	-0.115*	-0.373*	-0.122
Neighbourhood properties				
Non-missing 2007	0.29	-0.1	0.51*	0.12
Average property value neighbourhood	-0.00136*	0.00026	-0.00080	-0.00006
Non-missing 2005	0.11	-0.12	0.74*	1.04*
Percentage rented houses neighbourhood	-0.0073*	-0.0011	-0.0000	-0.0022
Percentage w.low income neighbourhood	0.0025	0.0068	-0.0207*	-0.023*
Maximum for average age	78.2	0[b]	10.9	25.8

[a] Reference category
[b] No age effect present

As for type of household, couples and single parent households are easier to contact than single person households. As these households are larger, the probability of finding at least one person at home is higher. After correcting for age etc., married couples without children and single parent households have a higher probability of agreeing to participate in the panel. The fraction of households that finally participate in the panel after having agreed to participate is higher for married couples. The fact that the number of additional adults in the household has a negative influence on primary and secondary participation, might be caused by the fact that adult children are less willing to participate in a household interview.

Households with first generation immigrant members are less easy to contact. As was mentioned earlier, this might be caused by language problems. Language problems during contact are marked as non-contact. There are no differences in contact and primary response between households with exclusively second generation immigrants and households with exclusively native Dutch members. The secondary response is relatively high for households with second generation non-western members. Quite a large fraction of households with first generation non-western members that have agreed to participate in the panel, finally do not participate. The reason for this last effect is unknown.

The influence of the average age of the adult household members confirms the results of the univariate analysis. Younger households are less easy to contact, as the response reaches a maximum for a relatively high age. However, younger households are more willing to participate in the internet panel, which is almost surely caused by the fact that younger households will more often already have a computer with internet access. The high primary response for older households suggests that older households are in principle willing to participate in surveys. The low response for these households for the internet panel, is therefore probably caused by the fact that these households do not have access to internet. The fact that they are offered a simple to use computer with internet access, does not completely solve this problem.

Households on addresses with more than one household are easier to contact. Also the conditional secondary and tertiary response is higher for these households. The higher contact could be caused by the fact that the probability of finding persons at the address when the interviewer visits is higher. As was mentioned in section 2.2, the households are partially imputed. While this gives correct totals, on a micro level there may be some errors. Most households on addresses with more than one household are imputed. The higher response for these addresses can therefore also be caused by wrongly imputed households. For example, two persons living at an address may be imputed as two single person households while they are in fact an unmarried couple. Since the response for unmarried couples is higher than that of single person households, this would result in an increase of the regression parameter for the number of households on the address.

4 Effect of offering a computer and internet access

People without internet access or computer cannot participate in internet surveys. This is usually a selective group, e.g. approximately 95% of the Dutch population aged 55

years or younger has internet access in 2008, while only 57% of the population aged 65 years or older has internet access (Statistics Netherlands, 2008*a*). In order to attenuate this effect, households without computer or internet access were offered a simple to use computer with internet access.

In order to investigate whether or not this has a positive effect on the representativity, the model for the tertiary response from section 3.3.1 was re-estimated, where we assumed that those households that receive the pc or internet access would not have responded. Table 6 shows the parameter values for these two models. At the bottom of the table are also shown the Nagelkerke pseudo R^2 values for the models.

The difference between the single person households and the other types of households (except 'Other/unknown') increases when the households that receive a computer with internet access from the MESS project are considered not to respond. Also the differences between households with first generation immigrants and those with solely second generation immigrants or native Dutch members increases. The largest influence is to be expected for the average age of the household members. The age at which the response reaches a maximum decreases only a little when the households without own computer are considered not to respond. However, the more negative value for the parameter for age squared indicates that the response decreases more quickly with age. This can also be seen in figure 10, where the response fraction is shown in black and dark grey and the non-response fraction in light grey for different age categories. The dark grey area are households that have received a computer from the MESS project. It is clearly visible that the fraction increases with age. Loss of these households from the panel has more influence for the older ages and will increase the differences in response for the different age categories.

Figure 11 shows the effect of offering a computer for the different types of household. We see especially an increase in response for single person households, married couples without children, and single parent households. The increase for the first two groups can be explained by the relatively high age of these groups. Figure 12 shows that offering internet access seems to have little influence on the representativity of the different ethnic groups. Only the response for immigrants from the Netherlands Antilles shows a relative increase.

The influence of offering a computer and internet on the representativity can also be seen in the change in the Nagelkerke R^2 value. As was discussed in section 2 the better the model can explain the response, the worse the representativity is. As the Nagelkerke R^2 increases when the households that receive the computer are considered not to respond, the representativity is higher when this group is added to the panel.

We can therefore conclude that offering of a simple to use computer with internet access has a positive influence on the representativity.

5 Comparison between LISS and the LFS

In the previous sections, the representativity of the response of the LISS-panel was investigated. However, almost every survey or panel suffers from underrepresentation of certain groups. As we have seen the LISS panel is no different. However, the

Table 6: Estimated parameters for logistic model for tertiary (overall) response. The first column shows the original parameter values (the same as in table 4). The second column shows the estimated values when households that borrow a computer or ADSL are considered not to respond. Values with a p-value below 0.05 are marked with ''.*

Parameter	Original	No pc/ADSL
Intercept	-2.21*	-2.48*
Household properties		
Household type=Single person	0[a]	0[a]
Household type=Unmar.couple w/o children	0.198*	0.304*
Household type=Mar.couple w/o children	0.569*	0.75*
Household type=Unmar.couple w.children	0.33*	0.44*
Household type=Mar.couple w.children	0.515*	0.678*
Household type=Single parent	0.31*	0.34*
Household type=Other/unknown	-1.28*	-1.14*
Ethn.=Native Dutch	0[a]	0[a]
Ethn.=First generation western	-0.48*	-0.48*
Ethn.=First+second generation western	-0.5*	-0.45*
Ethn.=Second generation western	-0.018	-0
Ethn.=First generation non-western	-0.83*	-0.93*
Ethn.=First+second generation non-western	-0.67*	-0.77*
Ethn.=Second generation non-western	0.29	0.25
Number of households at address	0.306*	0.3*
Average age	0.0704*	0.08*
Average age squared	-0.001094*	-0.00128*
Registered telephone	1.678*	1.583*
Number of additional adults	-0.282*	-0.273*
Neighbourhoood properties		
Non-missing 2007	0.45*	0.41*
Average property value neighbourhood	-0.00089*	-0.00072
Non-missing 2005	0.53*	0.78*
Percentage rented houses neighbourhood	-0.0042*	-0.004*
Percentage w.low income neighbourhood	-0.01*	-0.0161*
Maximum for average age	32.2	31.4
Nagelkerke R^2	0.31	0.33

[a] Reference category

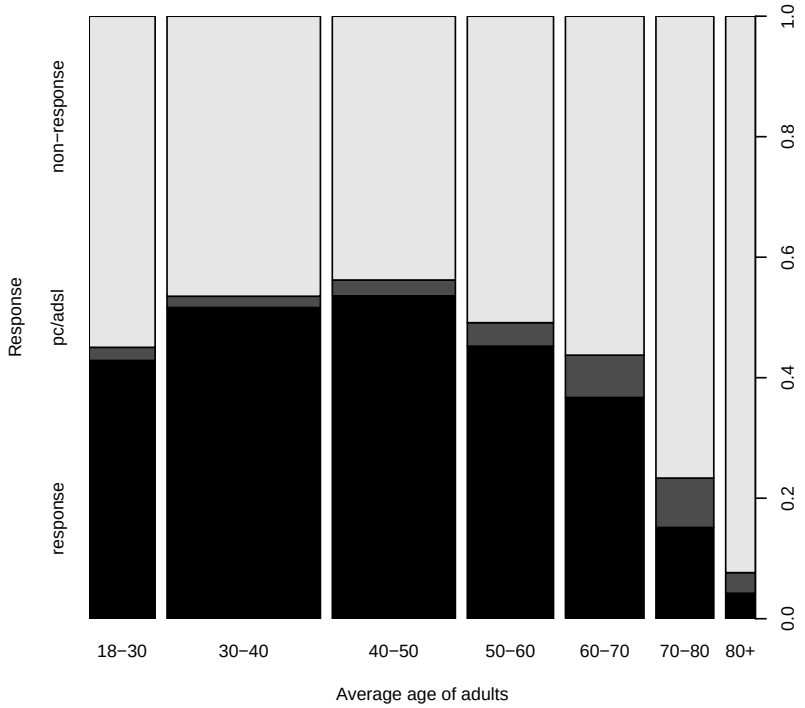


Figure 10: Effect of offering a computer and internet access on the response as a function of the average age of the household members of 18 years and older. In dark grey are shown the responding households that receive a computer and internet access during participation in the panel.

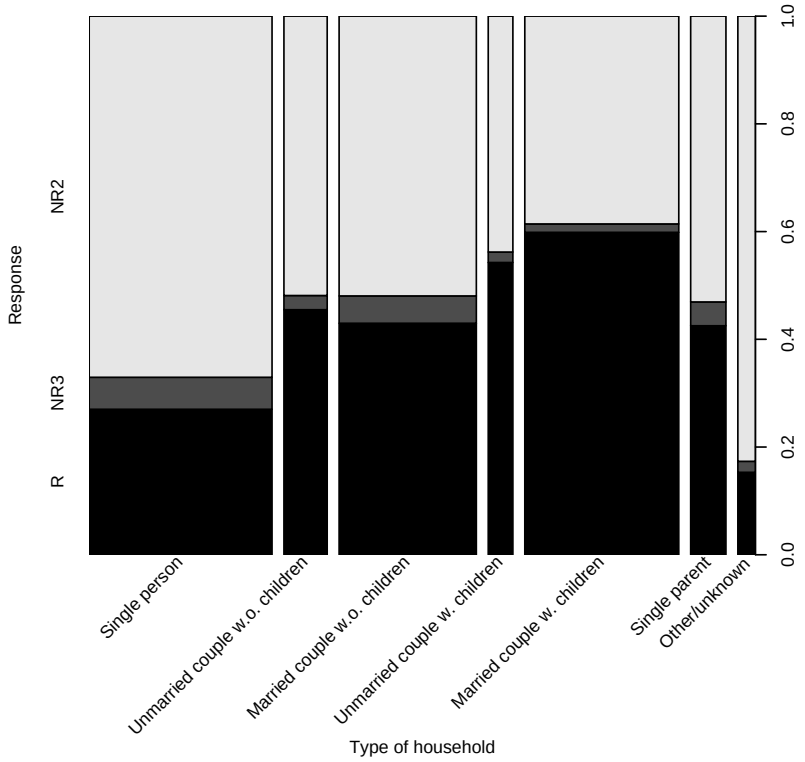


Figure 11: Effect of offering a computer and internet access on the response for the different types of households.

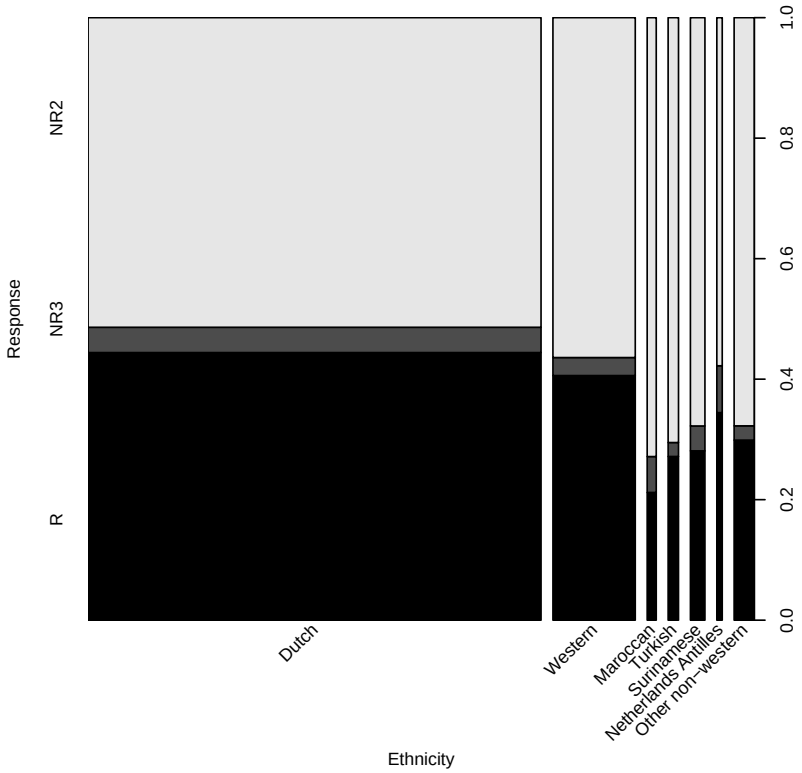


Figure 12: Effect of offering a computer and internet access on the response for the different ethnic groups.

costs involved in maintaining an internet panel are lower than that of a panel using interviewers. Therefore, a comparison between LISS panel and a more traditional non-internet panel would be very interesting. To do this, we have compared the LISS-panel to the (first wave of the) Dutch Labour Force Survey (LFS) of 2006 (see Statistics Netherlands, 2008*b*). The LFS is also a household survey and is one of the most important surveys of Statistics Netherlands. Therefore, it is a good candidate for a benchmark.

Like the LISS-panel, the LFS uses a random sample of addresses. However, unlike the LISS-panel all households on an address are first interviewed using an interviewer (CAPI). For the first interview no households are contacted using telephone. In this aspect the LFS differs from the LISS-panel where those households with a registered phone are initially contacted using telephone (CATI).

As the LFS focusses on the Dutch labour force, addresses with only persons aged 65 years or older are underrepresented. We therefore restricted our comparison to households with an average age of the adults below 65-years. In the LFS a household is considered to respond when all household members respond, but proxy-interviews are allowed for. Using the personal identification numbers present in the LFS response files the households were coupled to the household register and the same household properties were derived for the responding and non-responding households as in the investigation of the LISS-panel. The number of households used in the analyses for the LFS is 64,218 and for the LISS-panel 8,025.

Table 7: Respons rates for the LFS and LISS.

	LFS	LISS
Contact	84.7%	85.7%
Primary response	62.0%	69.4%
Tertiary response	-	45.9%

5.1 Response rates

Table 7 shows the response rates of the LFS and the LISS-panel[3]. The contact rates are very similar in both surveys. However, the primary response is slightly higher for the LISS-panel. The tertiary response cannot be compared, since only response data for the first interview of the LFS was available[4].

5.2 Univariate comparison

As was already mentioned, high response rates by themselves do not automatically imply a higher representativity and vice versa. We therefore also compared the LISS-panel to the LFS by comparing the response of different groups. As the overall response rates differ between the LFS and the LISS-panel, the response rates for different groups cannot be compared directly. Therefore, the response rates were normalised by dividing the response rate r_i^s of a group i by the response of a reference group r_0^s for each survey s:

$$\frac{r_i^s}{r_0^s} \tag{2}$$

We compared both the tertiary and primary response of the LISS-panel to that of the LFS. The comparison between the tertiary response and the LFS is of course the most interesting, but the primary response should be more similar to the LFS as it is a CAPI/CATI survey.

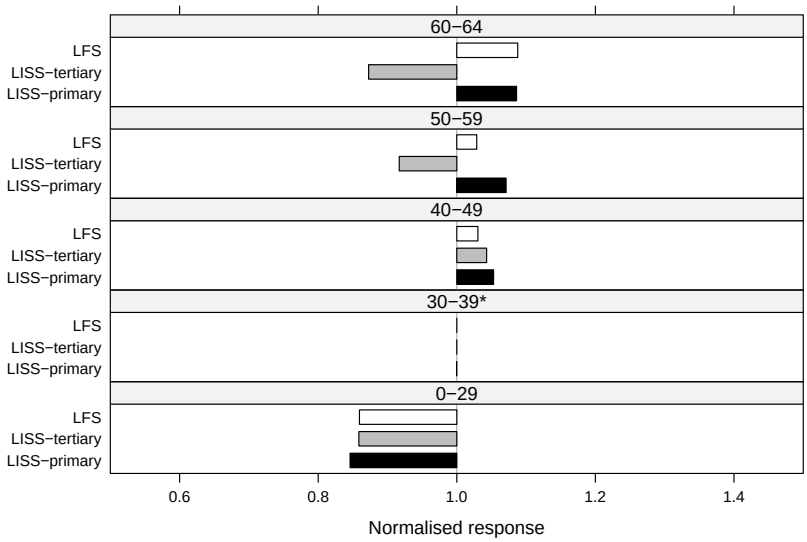
Figures 13 and 14 show the normalised response rates. When comparing the different age categories, it is clearly visible that although the primary response of the LISS-panel and the LFS are very similar, the tertiary response of the LISS-panel drops for higher age categories. Previous sections already showed that that older households are less inclined to participate in an internet panel.

The response of households with only second generation immigrants is higher than that of the LFS. However, the response of households with solely first generation immigrants is much lower. For the tertiary response this can probably be explained by the lower internet access in these groups, but this does not explain the differences between the primary response and the LFS.

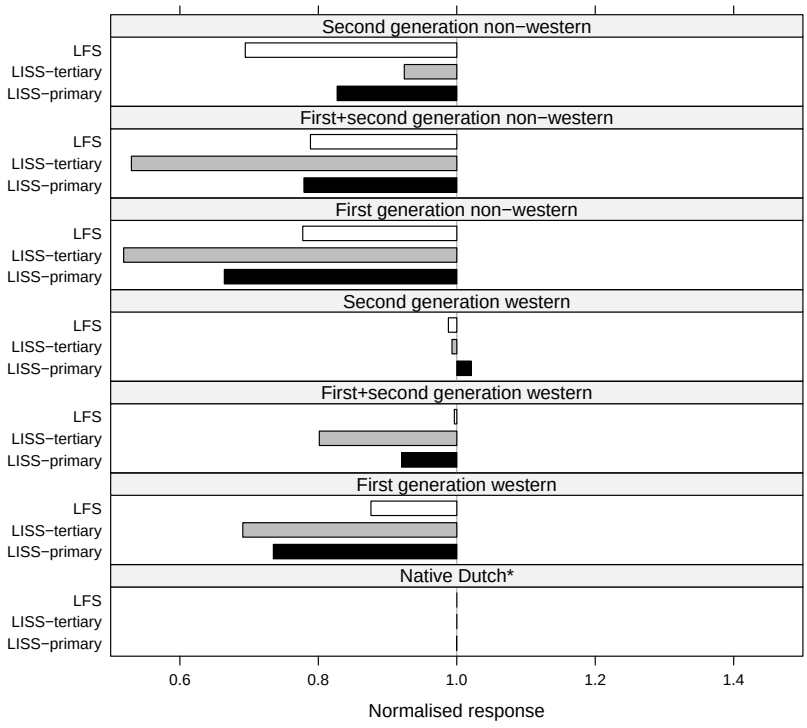
For the type of household, the primary response and the response of the LFS are again very similar. For the panel, especially the response of the single person households

[3]The response rates of the LISS-panel differ somewhat from the official rates published by CenterData see section 3.1.

[4]After the first interview people are contacted four times with a three month spacing using a telephone interview (CATI). However, information about these contact attempts was missing from the available response files.

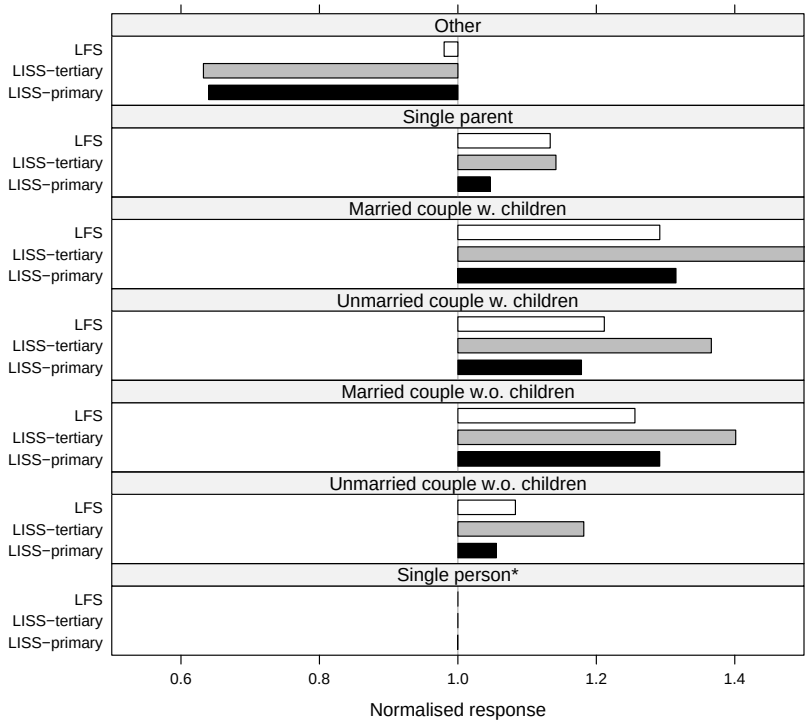


(a) Average age

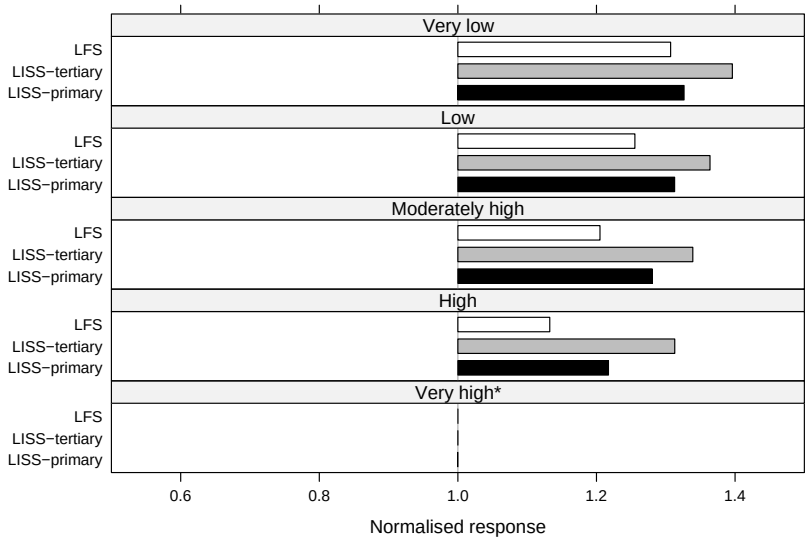


(b) Etnicity

Figure 13: Normalised response rates for the LFS, and the primary and tertiary response of the LISS-panel for different background variables. The reference categories are marked with ''.*



(a) Type of household



(b) Urbanisation

Figure 14: Normalised response rates for the LFS, and the primary and tertiary response of the LISS-panel for different background variables. The reference categories are marked with ''.*

is lower, which is partly caused by the higher age in this group. As for the amount of urbanisation, for both the primary and tertiary response, household in highly urbanised areas are underrepresented. This of course corresponds to the lower response of immigrants.

5.3 Multivariate comparison

For the multivariate analyses all data was combined into one data set. The data from the LISS-panel was duplicated, where the response variable of one set contains the primary response and the response variable of the other set contains the tertiary response. A categorical variable was created that indicated if the record belongs to the LFS, the primary response of the LISS-panel or the tertiary response of the LISS-panel. A logistic model was estimated using the common background properties. For each variable an interaction effect with the survey was added to the model. Therefore, effectively separate models were estimated for each of the surveys. However, differences between the LFS and the LISS-panel can more easily be seen, because the LFS was chosen as a reference category. A significant interaction effect indicates a significant difference from the LFS response. Table 8 shows the estimated model parameters.

The results confirm the results of the univariate analysis. The differences between the tertiary response and the LFS are generally larger than those between the primary response and the LFS. Married couples are relatively overrepresented for both primary and tertiary response. Compared to the LFS, households with only first generation immigrants are underrepresented, while household with only second generation non-western immigrants are overrepresented. The higher tertiary response of this last group might be explained by the relative young age of this group. However, this does not explain the higher primary response. As we have already seen, the tertiary response is lower for higher ages. The strong correlation between response and possession of a registered telephone is discussed below.

As we have already discussed in section 2.2, the better the model is able to predict the response, the bigger the differences in the individual response probabilities and the worse the representativity is for these variables. This can be expressed in a representativity indicator, where a value of one corresponds to a representative survey and lower values indicate increasing under- and over-representation of certain groups (Schouten and Cobben, 2007). When $\hat{p}_i$ is the response probability of respondent i predicted by the model, N the number of households in the sample and $\hat{\bar{p}}$ the estimated average response probability, the R-indicator R_1 is given by (Schouten and Cobben, 2007)

$$\hat{R}_1 = 1 - 2\sqrt{\hat{\text{var}}\{\hat{p}_i\}} = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{p}_i - \hat{\bar{p}}\right)^2}. \tag{3}$$

A value closer to one indicates a good representativity and a value close to zero worse representativity. The result of course depends on the variables that are taken into account in the model of the response. A more extensive discussion of R-indicators can be found in Schouten and Cobben (2007).

The indicator can also be used to estimate the maximum bias B_{max} that the non-representativity can introduce. The maximum bias occurs when we try to measure

Table 8: Coefficients of the response model. This model was estimated using a categorical variable indicating the survey. The interaction of this variable was taken with each of the variables in the model. The LFS was the reference category. Significant parameters in the last two columns indicate a significant difference from the LFS response. Significant parameters (p-value below 0.05) are marked with ''.*

Parameter	Reference	Primary LISS	Tertiary LISS
Intercept	−0.04	−0.14	−0.71
Household type=Single person	0[a]	0[a]	0[a]
Household type=Unmar.couple w/o children	0.197∗	−0.04	−0.02
Household type=Mar.couple w/o children	0.327∗	0.131	0.269∗
Household type=Unmar.couple w.children	0.307∗	0.15	−0.07
Household type=Mar.couple w.children	0.430∗	0.295∗	−0.038
Household type=Single parent	0.239∗	−0.00	−0.11
Household type=Other	0.159	−0.71∗	−0.59
Ethn.=Native Dutch	0[a]	0[a]	0[a]
Ethn.=First generation western	−0.276∗	−0.43∗	−0.31∗
Ethn.=First+second generation western	−0.069	−0.47	−0.46
Ethn.=Second generation western	−0.029	0.11	−0.008
Ethn.=First generation non-western	−0.435∗	−0.229∗	−0.41∗
Ethn.=First+second generation non-western	−0.459∗	−0.09	−0.56∗
Ethn.=Second generation non-western	−0.456∗	0.47∗	0.63∗
Average age	−0.0028	0.006	0.023
Average age squared	0.000139∗	−0.00009	−0.00060∗
Urbanisation=Very high	0[a]	0[a]	0[a]
Urbanisation=High	0.040	0.204∗	0.196∗
Urbanisation=Moderately high	0.099∗	0.154	0.000
Urbanisation=Low	0.118∗	0.120	−0.087
Urbanisation=Very low	0.240∗	−0.06	−0.29∗
Registered telephone	0.217∗	1.040∗	1.289∗
Number of children	0.110∗	−0.138∗	−0.003
Region=Northern Netherlands	0[a]	0[a]	0[a]
Region=Eastern Netherlands	−0.026	−0.14	−0.21∗
Region=Western Netherlands	−0.127∗	0.01	−0.049
Region=Southern Netherlands	0.131∗	−0.09	−0.14

[a] Reference category

Table 9: R-indicator (R_1), average response ($\hat{\bar{p}}$) and maximum bias (b_{max}) for a model with and without (*) 'registered phone' as covariate.

	$\hat{R}_1$	$\hat{R}_1^*$	$\hat{\bar{p}}$	b_{max}	b_{max}^*
LFS	0.81	0.82	0.619	0.16	0.15
LISS-primary	0.67	0.74	0.696	0.24	0.18
LISS-tertiary	0.58	0.75	0.515	0.40	0.25

a variable that is perfectly correlated with response. One such an example would be the question: 'Do you want to participate in our survey?' The maximum bias of course depends on the variance in the variable that one tries to estimate. Should everyone always give the same answer (variance equal to zero), it does not matter who one asks and the answer of respondents and non-respondents will be the same. If on the other hand there is a large spread in possible answers, then the differences between the respondents and non-respondents can also be larger. The maximum bias is given by (Schouten and Cobben, 2007).

$$B_{max} = \frac{1-\hat{R}_1}{2\hat{\bar{p}}}\sqrt{\text{var}\{Y\}} = b_{max}\cdot\sqrt{\text{var}\{Y\}}, \qquad \text{with } b_{max} \equiv \frac{1-\hat{R}_1}{2\hat{\bar{p}}}. \tag{4}$$

Table 9 shows $\hat{R}_1$ for the LFS, and the primary and tertiary response of the LISS-panel together with the average response probability and maximum bias. The R-indicator of the LFS is clearly much closer to one than those of the LISS-panel. This is reflected in the maximum bias, which is over 2.5 times larger for the tertiary response of the LISS-panel than that of the LFS. It should be noted that the comparison is only for households with an average age below 65 years. Should we have been able to also take older households into account in the comparison, the difference would probably have been larger, as the tertiary response drops for older households.

A further look into the reasons for this large difference in representativity showed that one of the most important causes is the fact that the number of contact attempts was too low for the households that were approached using interviewers (CAPI) in the LISS-panel. As we have mentioned earlier, households without registered telephone line were approached using CAPI, the households with a registered telephone were approached using CATI. Those households that could not be contacted using CATI were later approached using CAPI. The required number of contact attempts for CAPI was eight and for CATI fifteen.

Table 10 shows the response rates for the two initial modes. The contact rate of the households in the CAPI group is much lower than that of the households in the CATI group. This could of course be caused by the fact that households with a registered telephone are easier to contact. From the significant effect of registered telephone in table 8 for the LFS this indeed appears to be the case, and also in a pilot where a group of households with a registered telephone were contacted using CAPI, a similar effect was observed. However, the effect is much stronger in the LISS-panel than in the LFS. A closer look at the number of contact attempts showed that for 16% of the households in the CAPI group and 5% of the household in the CATI the number of contacts was less than the required eight and fifteen times. The plan is to complete the recruitment

Table 10: Response rates for each of the two initial methods of contact. The first two columns show to fraction of households participating in each of the steps. The last two columns show the response conditional on participation in the previous step.

	Total		Conditional	
	CATI	CAPI	CATI	CAPI
Contact	90.2%	75.6%	90.2%	75.6%
Primary response	74.3%	58.4%	82.3%	77.2%
Secondary response	54.0%	45.6%	72.7%	78.1%
Tertiary response	48.9%	39.2%	90.6%	85.9%

of these households in 2009.

The difference in the probability of contact between the CAPI and CATI, will probably also have an influence on other variables than 'registered telephone'. For example, immigrants and younger households more frequently have no registered telephone than respectively native Dutch or older households and are therefore more often contacted using CAPI.

When we leave the effect of 'registered telephone' out of the model as this is the variable that is most strongly related to the mode, we see in table 9 that the R-indicator of the LISS-panel is much closer to that of the LFS. This suggests that with enough contact attempts (enough to remove the mode effect), the representativity of the panel can be quite close to that of the LFS.

6 Conclusion

We investigated the representativity of the LISS internet panel, which is part of the MESS project. One of the main goals of this project is to set up a representative longitudinal panel of households. In order to obtain a representative panel, traditional sampling methods were used and households without computer or internet access were offered a simple to use computer with internet access. In order to investigate the representativity of this panel we compared the respondents and non-respondents to each other using background variables available for both the respondents and non-respondents.

When looking at the overall response, the response differed between the different types of households. Especially the response for single person households is lower than that of the other types of households. Furthermore, the response depends on the average age of the household members, the generations of immigrants present in the household, number of persons in the household, possession of a registered telephone, number of minors in the household, status of the neighbourhood, and the number of households living on the address of the household. The last effect is at least partially caused by the fact that some of the households are imputed.

Immigrants and people without internet access or computer are of especial interest in the MESS project, since these are usually underrepresented in internet panels. As for immigrants, for the overall response there were no significant differences between the

different ethnic backgrounds of the immigrants. The only effect that could be seen was that of the generation of immigrants. Households with first generation immigrants respond less than households with solely second generation immigrants or only native Dutch members. There were no differences between households with solely second generation immigrants and native Dutch members. This suggest that language problems play a role in the relatively high non-response of households with first generation immigrants. Which is also confirmed by the fact that especially in the contact phase a lot of immigrants are lost. It should be noted, however, that the fraction of first generation immigrants differs substantially between the different ethnic backgrounds. Some ethnic groups are therefore underrepresented.

In order to not have an underrepresentation of groups that do not have access to internet or a computer, these households are offered a simple to use computer with internet access that they can use for the duration of the panel. We investigated the effect of this on the representativity. Results show that the representativity is improved. Especially the representativity of households with older members is improved.

Almost every survey or panel is plagued by underrepresentation of certain groups. As we have seen the LISS panel is no different. However, the costs involved in maintaining an internet panel are lower than that of a panel using interviewers. Therefore, a comparison was made between the LISS panel and the LFS. This analysis confirmed the results found in the previous analyses: older households, households with first generation immigrants and single person households are underrepresented compared to the LFS. The multivariate analysis showed large differences in representativity between the LFS and the LISS panel. A further investigation showed that these differences are largely caused by differences in response for the two modes used in the recruitment of the LISS panel. When possession of a registered telephone line was removed from the model, the representativity of the LISS panel with regard to the variables included in the model was much closer to that of the LFS.

References

Bethlehem, J. (2006), 'Representativiteit van web-surveys — een illusie?', Nota, Centraal Bureau voor de Statistiek, Voorburg.

Harmsen, C. and Israëls, A. (2003), Register based household statistics, *in* 'European population conference 2003'.

Schouten, B. and Cobben, F. (2007), 'R-indexes for the comparison of different fieldwork stategies and data collection modes', *Discussion papers* **07002**, pp. 1–29.

Statistics Netherlands (2005/2007), 'Regional key figures of the Netherlands'. `http://statline.cbs.nl/StatWeb/selection/?DM=SLEN&PA=70072ENG&LA=EN&VW=T`.

Statistics Netherlands (2008*a*), 'ICT gebruik van personen naar persoonskenmerken'. `http://statline.cbs.nl/StatWeb/selection/?VW=T&DM=SLNL&PA=71098ned&D1=0-2,33-133&D2=0-2&D3=a&HDR=G1&STB=T,G2`.

Statistics Netherlands (2008*b*), 'Methoden en definities Enquête Beroepsbevolking 2007'. http://www.cbs.nl/nl-NL/menu/methoden/dataverzameling/onderzoeksbeschrijving-ebb-art.htm.