Discussion paper

Decomposition of mode-specific measurement biases and selection biases in health statistics

Barry Schouten
Kees van Berkel
Hamza Ahmadan
Nino Mushkudiani

October 2024

Content

1. Introduction 4

2. Re-interview design 6
   - 2.1 General considerations 6
   - 2.2 Categorizing health survey statistics 7
   - 2.3 Selection of strata at-risk 9
   - 2.4 A re-interview in the Dutch Health Survey 11
   - 2.5 Ensuring re-interview assumptions 13

3. Estimation strategy 13
   - 3.1 Estimating mode-specific measurement bias for web respondents 14
   - 3.2 Adjustments for structural time effects 15

4. Results 15
   - 4.1 Re-interview fieldwork and assumptions 15
   - 4.2 What proportion of mode-specific biases can be attributed to measurement? 18
   - 4.3 Are the directions of mode-specific measurement biases in line with expectations? 19

5. Discussion 21

References 24

Appendix A - Interviewer training 26

Appendix B - Auxiliary variables linked from administrative data 27

Summary

Mixed-mode survey designs and adaptive survey designs have become a standard in many survey settings. Also the combination of mixing modes and adapting effort has been elaborated and applied. However, the focus has been on balancing cost and nonresponse; measurement differences between modes are mostly ignored. Measurement error estimates require sophisticated, additional data collection that is costly. In this paper, we estimate mode-specific measurement biases between web and face-to-face for a range of health statistics. We show that mode-specific measurement biases can be large and cannot be ignored. This has implications for both mixed-mode survey design and for adapting effort in such designs.

Keywords

Mode effects, Mixed-mode surveys, Adaptive survey design, Health Survey

1. Introduction

In this paper, we provide estimates for mode-specific measurement and mode-specific selection biases for a wide range of health statistics between web and face-to-face interviews. The statistics are derived from the Dutch Health survey and are also part of the European Health Interview Survey (EHIS) that is conducted every six years within the European Statistical System (ESS).

To date, mixed-mode survey designs are omnipresent and a wide array of literature describes the consequences for design and analysis (e.g. Dillman, Smyth & Christian 2014, Roberts & Vandenplas 2017, Sakshaug, Cernat & Raghunathan 2019, Yu, Elliott & Raghunathan 2024). In parallel, also the methodology for adapting effort in survey design, so-called adaptive survey designs, has been elaborated and described (e.g. Schouten, Peytchev and Wagner 2017). However, adaptive survey designs have been predominantly focused at minimizing nonresponse/selection biases and mostly ignored the impact of measurement error. Being the most influential design feature in terms of response and cost, but also in measurement, adapting effort through modes must acknowledge both survey errors. While some attempts have been made (e.g. Calinescu 2013), literature still is very thin. One likely reason for this is the complexity of estimating the differential impact of design-specific features on measurement. It is only through advanced designs that such estimates can be obtained. Another likely reason is that differential measurement error has a much more implicit relation to cost. And controlling cost is often a driving force behind going adaptive. A key prerequisite and first step towards adaptive mixed-mode survey designs is accurate estimates per stratum of mode-specific measurement biases.

Statistics Netherlands adopted adaptive mixed-mode survey design as a standard for its person and household surveys. The survey that was migrated first to such a design was the Dutch Health Survey in 2018 (Van Berkel, Van der Doef & Schouten 2020). This survey employs a sequential mixed-mode design where face-to-face follows web. Web nonrespondents that did not explicitly refuse to participate are allocated to face-to-face interviewers. The allocation fraction differs over population strata and depends on stratum web response rates and anticipated stratum response rates to face-to-face. The focus is on balancing response rates across strata; any measurement effects between the two modes are ignored. However, the mode effects between the two modes are large, up to 10 percentage points for some of the health statistics. This difference in web and face-to-face estimates is the result of different types of respondents being recruited by interviewers and different answers given per mode. How large the two components are is unknown, but all overall mode effects point in the same direction: Estimates under face-to-face show a more negative health profile. This finding is contradictory to social desirability, where the expectation would be that respondents in face-to-face interviews provide more positive answers (Tourangeau & Yan 2007, Kreuter, Presser and Tourangeau 2008). From this finding comes the conjecture that mode-specific selection biases and mode-specific measurement biases between web and face-to-face may have opposite signs. They thus cancel

each other out to some extent. In addition, the mode effects are particularly large for questions that require recall and/or calculations. Examples are the average amount of sedentary behaviour in an arbitrary week and the average fruit consumption in an average week. This would point at a role for interviewers in assisting respondents and/or in keeping them motivated (Van der Vaart, Van der Zouwen and Dijkstra 1995, Van der Zouwen 2000). If these conjectures are true, then this would be crucial input to the interpretation of the health statistics and to the application of adaptive survey design in a mixed-mode setting.

Ultimately, we like to answer three research questions: 1) What proportion of mode-specific biases can be attributed to measurement?, 2) Are the directions of mode-specific measurement biases in line with social desirability?, and 3) Do the directions of mode-specific measurement biases point at interviewer effects in questions that require more cognitive effort?

In estimating biases between modes we employ a re-interview experimental design and follow the estimation strategy proposed by Klausch et al. (2017). Web respondents are invited for a face-to-face interview with the same questionnaire. The re-interview response is calibrated to the web response. The differences between the weighted and unweighted re-interview responses for a range of health statistics are treated as estimates for the mode-specific selection biases. The remaining differences between the web responses and the weighted re-interview responses then are viewed as estimates for the mode-specific measurement biases of the selected health statistics. In order to control for time change in health statistics, another sample of web respondents is approached once more with the same web questionnaire.

Obviously, the resulting estimates for mode-specific selection and measurement biases should be treated with care as they make assumptions that need not be valid. Under the assumptions, we find mode-specific measurement biases in the expected ‘social desirability’ direction. The selection biases are even larger than the total mode biases as they in part are compensated by measurement biases. We also find that questions requiring more cognitive effort experience a downward interviewer effect, i.e. point at overreporting of more ‘healthy’ behaviour in self-administered modes. These findings imply a complicated trade-off in adaptive survey design that is briefly discussed.

The remainder of this paper is divided into four sections. In Section 2, we present the re-interview design and selections we have made in the re-interview. We then explain the estimation strategy in Section 3. Our results are shown in Section 4. We discuss the consequences of the results in Section 5.

2. Re-interview design

In this section, we describe the re-interview experiment as carried out between August 2022 and March 2023 linked to the Dutch Health Survey. We, first, go more abstract on the utility of re-interview designs and discuss the assumptions underlying to such designs. Next, we motivate the selection of specific population strata that are anticipated to be at higher risk of mode-specific measurement biases. Finally, we present the design as it was implemented. Section 3 continues with the estimation strategy.

2.1 General considerations

Our interest is in accounting for mode-specific measurement biases in adaptive survey designs where the survey mode is a design feature. More specifically, we consider sequential mixed-mode survey designs where interviewer modes follow self-administered modes. This is the default data collection strategy at Statistics Netherlands, but such designs are applied as well in many other countries, mostly for costs efficiency. An adaptive interviewer follow-up to nonresponse in self-administered modes may be very attractive from a representation perspective. However, measurement differences between the modes may lead to incomparability in time when the shares of the modes shift and incomparability between relevant subpopulations when their shares in modes vary greatly.

Estimating measurement differences between modes is a renowned complex challenge. The main reasons are that measurement is confounded with selection and measurement is observed for one mode only. So there is one answer and the mode of measurement can only partially be controlled. A solution suggested in the literature (Schouten et al. 2024) is a re-interview design. In a two mode design, the re-interview looks like Figure 1. In Klausch et al. (2017) a design is presented for surveys with three different modes. In Figure 1, mode m_1 represents the first mode in the sequential mixed-mode design, i.e. a self-administered mode in our setting. Mode m_2 then is the follow-up mode, i.e. here an interview mode. Given that the design is sequential, there is a time lag between the two modes. At Statistics Netherlands, this time lag in between is one month; if a sample unit is a nonrespondent in month t than it is allocated to an interviewer in month t+2. Area A represents respondents to the first mode and area E respondents to the second mode. Nonrespondents to both modes, i.e. neither in A or E, are not shown as these do not contribute to mode-specific measurement differences. The respondents to the first mode are assigned to a re-interview with the second mode. Area C responds and area D does not. Consequently, areas B, the hypothetical answers in the first mode for nonrespondents to that mode, and D, the hypothetical answers in the second mode for re-interview nonresponse, are missing. In Section 3, it is explained how these are estimated. In practice, the respondents in A may be subsampled for the re-interview. The re-interview inclusion probabilities then need to be accounted for in the estimation.

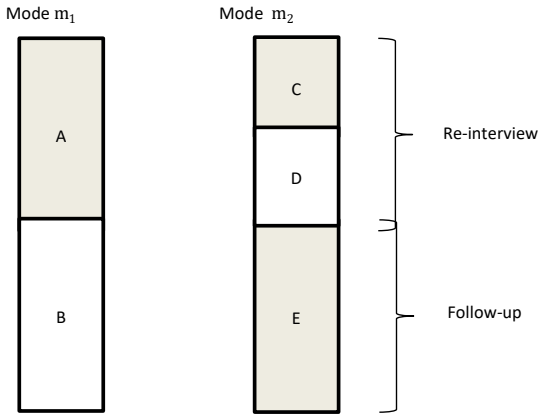

Figure 1: Re-interview design for a $m_1 \rightarrow m_2$ sequential survey design. Grey areas represent m_1 response (A), re-interview m_2 response (C) and follow-up m_2 response (E). White areas represent m_1 nonresponse (B) and re-interview m_2 nonresponse (D). Nonresponse to both modes is omitted.

It is important to stress that re-interview-based estimation of mode-specific measurement biases goes with assumptions. In order to be effective, re-interview designs require three assumptions:

1. Re-interview measurement behaviour is not affected by the first interview.
2. Re-interview nonresponse is preserving the relative measurement errors between the modes.
3. True values of the survey variables of interest have not changed between the first interview and the re-interview.

The first two assumptions may be combined into one assumption: The true relative measurement errors between the modes for the hypothetical answers without re-interview hold for the re-interview sample. In more simple terms, it is assumed that the measurement error model that is posed is unaffected by the re-interview. We will return to how we have tried to safeguard these assumptions for the Health Survey re-interview experiment in Section 2.5.

2.2 Categorizing health survey statistics

The Health Survey is a multi-purpose survey. It covers a wide range of topics such as general perceptions on health to physical, mental barriers, frequency of dentist and doctor visits, hospitalization, sports, nutrition, medicine use, and alcohol and drug use. The number of key statistics is broad and, consequently, also diverse in question characteristics. As such it is an ideal questionnaire to evaluate the impact of mode on measurement.

In order to evaluate biases, we divide the key Health Survey questions into four types:

1. OBJECTIVE – NON-SENSITIVE – EASY: These questions are anticipated to be at very small risk of mode-specific measurement bias.
2. OBJECTIVE – NON-SENSITIVE – HARD: These questions may be subject to recall and/or reporting bias. They are not subject to misunderstanding and are not prone to social desirability bias.
3. OBJECTIVE – SENSITIVE – EASY: These questions are prone to social desirability bias, but not misunderstanding or miscalculations.

4. SUBJECTIVE – SENSITIVE – EASY: These questions are prone to social desirability bias and simultaneously to differences in interpretation/understanding.

We label the four types as Robust, Complex, Sensitive and Subjective. We realize our 'taxonomy' and subsequent classification of questions is likely subjective. We tried to follow suggestions in literature on sensitive questions (Saris and Gallhofer 2007, Tourangeau and Yan 2007, Campanelli et al. 2011), on questions (potentially) arousing emotional response (Lensvelt-Mulders 2008), and on questions requiring recall/more cognitive effort (Van der Zouwen 2000, Beukenhorst et al. 2013). Given that we consider multiple questions per type, we believe the findings will be less sensitive to our own categorization. Table 1 displays 19 Health Survey statistics classified to the four categories. Two of the sensitive questions, 'Heavy drinkers' and 'Use of cannabis' are based on computer-assisted-self-interviewing (CASI), also in F2F interviews. Data are used from the year 2017 as this was the last year without any form of adaptive survey design. More recent years all have a form of targeted interviewer follow-up.

Table 1: Observed mode response means in GEZO 2017 for web and F2F respondents for key statistics of different question types. Standard errors are between 0.4 and 0.7%. 'Sensitive' questions 'Heavy drinker' and 'Ever used drugs' are based on computer-assisted-self-interviewing (CASI).

Type	Statistic/question	Response means (unadjusted, in %)	
		Web	F2F
Robust	Contact dentist (last 12 months)	82	77
	Contact GP (last 12 months)	68	69
	Contact physio/exercise therapist (last 12 months)	27	24
	Diabetes type 2	4	4
	Weekly sporter	59	50
Complex	Sufficient fruit during 7 days a week	31	34
	Sufficient vegetables during 7 days a week	21	23
	Sufficiently active at moderate intensity	54	37
	Sufficient balance exercises	19	20
	Sufficient muscle-bone strengthening activities	84	73
	Use of non-prescribed medicine(s) (last 14 days)	36	45
Sensitive	Smokers	16	26
	Severe overweight	12	11
	Heavy drinkers	8	9
	Use of cannabis (last year)	17	22
	Use of prescribed medicine(s) (last 14 days)	43	40
Subjective	Self-perceived health (very) good	79	81
	Psychological distress (MHI-5<60)	11	9
	Persons with ≥ 1 OECD limitation	30	26

Table 1 presents the response means of web and face-to-face (F2F) estimates for a number of Health statistics. Note that F2F is sequential to web, i.e. consist of web nonrespondents that were interviewed by a face-to-face interviewer. The general tendency is that F2F gives a more negative picture in (almost) all statistics. Given

that social desirability in F2F would point at a more positive picture, we conjectured beforehand that selection and measurement biases may have opposite signs, i.e. they cancel each other in part. For the 'robust' statistics, we expected to find mostly selection differences.

2.3 Selection of strata at-risk

A full re-interview of all nonresponse with interviewer modes while maintaining statistical power at the level of adaptive survey design strata is expensive. For this reason, we have chosen to focus attention to those subpopulations where mode-specific measurement biases may be the largest.

In order to select strata for the re-interview, we needed a criterion. For this we chose the actual coverage of confidence intervals for a Health Survey statistic when the statistic is subject to measurement bias. Given that the intervals are specific to a statistic, we decided to base selection on variable Yes/no smoking and check strata selection on three other variables: yes/no obese, yes/no self-reported psychological issues and yes/no use of non-prescribed medicine.
We consider population strata to be eligible for the re-interview when the actual coverage of the 95% confidence interval for the smoking rate is smaller than a pre-defined threshold,

$$P\left(I_S^{95} \ni \overline{Y}_s\right) \leq \rho,$$

where I_S^{95} is the 95% confidence interval for the smoking rate $\overline{Y}_s$ in stratum S and ρ is the pre-defined threshold. For any stratum S, the response mean $\bar{y}_S$ of the smoking variable Y_s has a normal distribution in limit,

$$\bar{y}_S \sim N(\overline{Y}_s + B(\bar{y}_S), S(\bar{y}_S)),$$

where B is the bias and S the standard error. The bias is estimated by the absolute difference between the measurement of the fraction of smokers in a stratum based on web (WEB) and that based on F2F (F2F),

$$\hat{B}(\bar{y}_S) = \left| \frac{n_{cawi}^{smoking}}{n_{cawi}} - \frac{n_{capi}^{smoking}}{n_{capi}} \right|,$$

which implicitly assumes that either web or F2F is a measurement benchmark, i.e. is free of measurement error.

The stratum standard error, $S(\bar{y}_S)$, equals the square root of the sum of the response variances for the two modes, which we estimated on historic survey data.
The actual coverage of the 95% confidence interval I_S^{95} for $\bar{y}_S$ equals

$$P\left(I_S^{95} \ni \overline{Y}_s\right) = \Phi\left(1.96 - \frac{B(\bar{y}_S)}{S(\bar{y}_S)}\right) - \Phi\left(-1.96 - \frac{B(\bar{y}_S)}{S(\bar{y}_S)}\right),$$

where Φ denotes the cumulative distribution function of the standard normal distribution. We decided to take a coverage of 48% as a threshold, because then in the majority of cases the interval would be misplaced. This means that the estimated ratio of the measurement difference between smoking in web and smoking in F2F and the stratum standard error, estimated in the sample, must be greater than two in order to be selected for the re-interview.

A classification tree was used to divide the sample into increasingly smaller strata. The classification stopped when no more values above the threshold were found. To make sure that strata are not too small, a lower limit of 600 respondents was taken for each stratum. The auxiliary variables we included are age, ethnicity, marital status, income and urbanity. The following strata were found:

1. Migrants in all but non-urban areas.
2. Dutch people, aged 12-24 years, with income percentiles 20-100, never been married, in moderate to very high urban areas.
3. Dutch people, aged 25 years or over, with income percentiles 40-100, married or never been married, in very high urban areas.
4. Dutch people, aged 45 years or over, with income percentiles 40-100, married, in low urban areas.
5. Dutch people, aged 45-64 years, with income percentiles 40-100, married, in high or moderate urban areas.

Table 2 contains the results of the classification tree. The population distribution of the strata is also given. The first stratum is the largest and makes up around a quarter of regular samples. It has a value just above the selection threshold. Stratum 2 shows the largest ratio of 4.5, but makes up only around 6% of regular samples. Taking the strata together, the re-interview is applied to a subset making around 49% regular samples.

Table 2: Measurement differences (Diff), standard errors (SE) and ratios (Bias/SE) for the smoking rates per stratum. Also given are population distributions of the strata.

Stratum	Diff (in %)	SE (in %)	Bias/SE	Proportion (in %)
1	5.1	2.2	2.3	25.8
2	13.7	3.0	4.5	6.2
3	11.5	3.5	3.3	6.2
4	11.9	3.2	3.8	5.0
5	13.0	3.4	3.9	5.4

Table 3: Re-interview criterion applied to Non-prescribed Medicine, Overweight and Psychological distress.

Stratum	Non-prescribed medicine	Overweight	Psychological distress
1	4.59	0.38	3.20
2	3.42	0.78	1.88
3	1.84	1.50	0.89
4	0.99	0.89	0.38
5	2.29	0.35	1.07

We evaluated this classification by looking at the other key Health Survey variables. We recalculated the values $\frac{B(\bar{y}_s)}{S(\bar{y}_s)}$ for 'Non-prescribed Medicine', 'Overweight' and 'Psychological distress'. Table 3 contains the results for these three statistics. On Non-prescribed medicine threshold values are large. For Overweight they tend to be small. For Psychological distress only stratum 1 has a large ratio.

2.4 A re-interview in the Dutch Health Survey

We describe the design of the re-interview experiment. It had two experimental arms; one where web respondents were interviewed once more in F2F and one where web respondents were invited to once more do the web survey. The overarching sample was randomly split into the two arms.

Questionnaire design: The questionnaire used for the experiment is identical to that of the regular survey. As a result, there is no confounding of the questionnaire content with response and answer behaviour in the re-interviews.

Data collection strategy: Initially, two samples were selected to be approached through Web in August and September, with follow-ups conducted in October and November. The response to the re-interviews turned out to be lower than expected. Therefore, an additional sample was drawn and observed via Web in December, with re-interviews scheduled for February. Notably, March was added to the F2F observation due to a shortage of interviewer capacity in February. The observation strategy is shown in Figure 2. The selection of groups A and B for re-interview and group C for no follow-up is explained in the sampling design below.

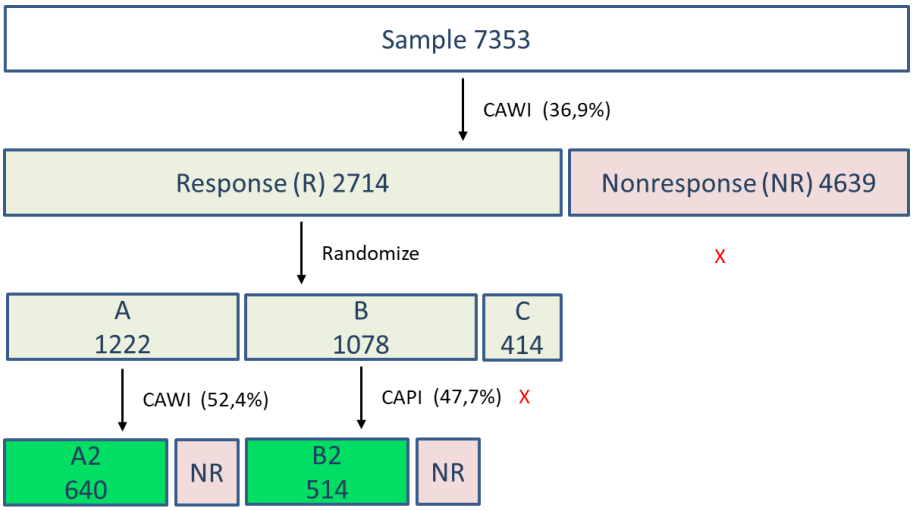

Figure 2: Data collection strategy re-interview experiment

The announcement of the re-interviews was made by letter. In the letter, the respondent is thanked for participating in the survey a few weeks ago. It is mentioned that a small number of people are being asked to participate in the survey again to gain further insights into the health of the population. The letter

for the F2F re-interview reminds the respondent that he/she completed the survey online last time but now a Statistics Netherlands employee will visit for a personal interview.

Sampling design: The initial sample for web observation is a stratified sample with sizes per stratum as listed in column Sample of Table 4. The Response column contains the numbers of people who completed the questionnaire online. Within each stratum, these respondents are divided into three groups: a group for web re-interview (A), a group for face-to-face (F2F) re-interview (B), and a group without re-interview (C). The response numbers for groups A and B are shown in columns A2 and B2, respectively.

Prior to setting the sampling design, a power analysis was performed. The type I error (false acceptance of a relative measurement difference) and the type II error (false acceptance of no relative measurement difference) were set at, respectively, 5% and 20% for a minimal observable relative measurement difference of 5%.

Table 4: Samples and responses per stratum

Stratum	Sample	Response	A	A2	B	B2	C
1	2147	524	257	109	267	108	0
2	1639	585	315	137	231	101	39
3	1147	500	205	123	183	105	112
4	1075	519	233	155	177	97	109
5	1345	586	212	116	220	103	154
Total	7353	2714	1222	640	1078	514	414

Interviewer instructions: Prior to the experiment, the participating interviewers received a briefing on its details, with a specific focus on its purpose and design. The briefing also covered potential pitfalls associated with completing a questionnaire, both online and face-to-face. When approaching people for re-interview, interviewers were instructed to communicate that the survey's objective is to depict the development of health characteristics. Interviewers appreciated the additional training sessions as they feared getting negative response at the doorstep. We have to note that the re-interview was relatively short after the COVID-19 pandemic. In Appendix A, we present a Q&A that was used in training.

Control group for structural time effects: As depicted in Figure 2, a control group A was added. These web respondents were once more invited to participate in web. The timing of the web re-interview was identical to that of the F2F re-interview group: at the first Friday of the second upcoming month. So there was always at least a time lag of four weeks, but for web respondents that were early respondents the invitation would be almost eight weeks later.

The control group had two purposes. The first purpose was to control for any time change in key health statistics. Although the time lag was between four and eight weeks, there could be seasonal influences or unexpected shocks. Health statistics are, however, relatively stable across time apart from seasonal influences on some of the statistics. The second purpose was to evaluate potential experimental

impact. This could be a true change in behavior caused by being interviewed and made more aware of lifestyle and health determinants. This could also be spurious change due to insufficient reliability, i.e. a respondent rethinking an earlier answer. The experimental design does not allow to disentangle structural time change from true experimental impact or spurious experimental impact, but the compound effect can be evaluated. When the estimated time change is small, then this gives more empirical ground to believe that re-interview assumptions, discussed in Section 2.5, hold for the F2F arm of the experiment.

The time change estimates are incorporated in estimates for relative measurement bias when they are large across multiple statistics. We present and discuss estimates in Section 4.4.

2.5 Ensuring re-interview assumptions

Re-interview designs essentially assume absence of experimental impact on measurement properties during the re-interview. As described in previous subsections, we have implemented a number of measures to prevent or detect such impact:

- The re-interview questionnaire is an exact copy of the original questionnaire.
- Interviewers received additional training based on anticipated questions at the doorstep.
- The re-interview timing is between four weeks and eight weeks which seems optimal in balancing respondent recall of the first mode answers and real change in Health Survey variables.
- Auxiliary variables from administrative data are linked to assist in adjusting potentially selective nonresponse in the re-interview.
- A control group was added to check for experimental impact and structural time change.

Nonetheless, it must be expected that re-interview assumptions are violated to some extent. For this reason, especially the overall impression across multiple health survey variables and the clustering of the variables in different types will be crucial.

One additional option to evaluate accuracy of estimates for mode-specific measurement biases is to consider mode share fluctuations over time. This is not done in this paper, but will be discussed in the last section to this paper.

3. Estimation strategy

Going from re-interview data to estimates for mode-specific measurement biases requires a number of steps. These are explained in this section. Throughout, for

convenience of exposition, we consider face-to-face data collection as the measurement benchmark.

3.1 Estimating mode-specific measurement bias for web respondents

The mode-specific measurement bias is estimated in six steps:

1. Model the nonresponse to the F2F re-interview using linked auxiliary variables and the Health Survey variables measured in web.
2. Per Health Survey variable:
   a. Model the Health Survey variable in the F2F re-interview using linked auxiliary variables and the Health Survey variables measured in web.
   b. Identify the variables selected in step 1 and/or step 2-a.
   c. Perform Generalized Regression (GREG) weighting to the F2F re-interview using the variables in step 2-b.
   d. Subtract the calibrated re-interview statistic from the statistic based on web response.
3. Apply a bootstrap sampling procedure to estimate standard errors/confidence intervals.

Step 2-d gives the estimated mode-specific measurement biases for web respondents.

We describe the estimator for the mode-specific measurement bias more explicitly. Let

$y_{web,web}^{R1}$ = mean of survey variable y of the respondents from the web interview selected for the web re-interview (A in Figure 2).

$y_{web,f2f}^{R1}$ = mean of survey variable y of the respondents from the web interview selected for the F2F re-interview (B in Figure 2).

$y_{web,web}^{R2}$ = mean of survey variable y of the respondents in the web re-interview (A2 in Figure 2).

$y_{web,f2f}^{R2}$ = mean of survey variable y of the respondents in the F2F re-interview (B2 in Figure 2).

$y_{web,web}^{R2cal}$ = calibrated mean of survey variable y of the respondents in the web re-interview (A2 in Figure 2), where the calibration is towards the web respondents (A in Figure 2).

$y_{web,f2f}^{R2cal}$ = calibrated mean of survey variable y of the respondents in the F2F re-interview (B2 in Figure 2), where the calibration is towards the web respondents (B in Figure 2).

The estimated relative measurement bias of F2F then is $\left(y_{web,f2f}^{R2cal} - y_{web,f2f}^{R1}\right)$.

3.2 Adjustments for structural time effects

Analogously, the time effect based on the web re-interview is also estimated in six steps:

1. Model the nonresponse to the web re-interview using linked auxiliary variables and the Health Survey variables measured in web.
2. Per Health Survey variable:
   e. Model the Health Survey variable in the web re-interview using linked auxiliary variables and the Health Survey variables measured in the first web measurement.
   f. Identify the variables selected in step 1 and/or step 2-a.
   g. Perform GREG weighting to the web re-interview using the variables in step 2-b.
   h. Subtract the calibrated re-interview statistic from the statistic based on the first measurement web response.
3. Apply a bootstrap sampling procedure to estimate standard errors/confidence intervals.

Depending on the size and significance of the estimated time effects, the estimates can be subtracted from the estimates in Section 3.1. If time effects are small, we ignore them in order to not inflate standard errors. To estimate time effects standard errors, the bootstrap sampling loop must encompass both calibration steps.

The estimator for the time effect is $\left(y_{web,web}^{R2cal} - y_{web,web}^{R1}\right)$, and the estimator for the relative measurement bias adjusted for the time effect is $\left(y_{web,f2f}^{R2cal} - y_{web,f2f}^{R1}\right) - \left(y_{web,web}^{R2cal} - y_{web,web}^{R1}\right)$.

The above strategy does not yet 'mix' the measurement-bias-adjusted estimates of web respondents with the estimates of the F2F follow of web nonresponse. Hence, as a last step, the full mixed-mode estimate is derived by weighting the adjusted web estimates and the F2F estimates according to their proportional size in the total response.

4. Results

In this section we answer the three research questions, stated in the Introduction. We start, with a description of the fieldwork and an evaluation of re-interview assumptions.

4.1 Re-interview fieldwork and assumptions

Before looking at mode-specific measurement bias estimates, we need to consider the validity of underlying assumptions. The re-interview estimates lean on the

absence of impact of the first interview on the second interview. Furthermore, we assume that real changes in Health Survey statistics can be captured by the web re-interview sample. We, therefore, first evaluate the response to both arms in the re-interview. Next, we look at the structural change estimates in the web re-interview arm. We must, however, note that we cannot directly evaluate validity without knowing 'true values' of the statistics. We can only find indirect evidence.

Table 5: Cramer's V for various variables against F2F re-interview response and against web re-interview response.

Variable	Re-interview response	
	F2F	Web
Contact dentist (last 12 months)	0.005	0.004
Contact GP (last 12 months)	0.024	0.054
Contact physio/exercise therapist (last 12 months)	0.009	0.046
Diabetes type 2	0.026	0.022
Weekly sporter	0.056	0.017
Sufficient fruit during 7 days a week	0.029	0.055
Sufficient vegetables during 7 days a week	0.022	0.048
Sufficiently active at moderate intensity	0.044	0.022
Sufficient balance exercises	0.058	0.104
Sufficient muscle-bone strengthening activities	0.034	0.002
Use of non-prescribed medicine(s) (last 14 days)	0.048	0.049
Smokers	0.069	0.049
Severe overweight	0.031	0.022
Heavy drinkers	0.021	0.026
Use of cannabis (last year)	0.079	0.069
Use of prescribed medicine(s) (last 14 days)	0.030	0.106
Self-perceived health (very) good	0.010	0.025
Psychological distress (MHI-5<60)	0.072	0.022
Persons with ≥ 1 OECD limitation	0.008	0.039
Type of household	0.134	0.239
Position in household	0.119	0.244
Marital status	0.098	0.184
Age	0.115	0.257
Gender	0.021	0.052
Migration background	0.100	0.157
Socio-economic status (SES)	0.118	0.245
Household income (quintiles)	0.088	0.092
Total assets (quintiles)	0.079	0.109
House ownership	0.099	0.079
Province+	0.111	0.118
Region	0.046	0.050
Urbanization degree	0.086	0.126

The F2F re-interview had a response rate of 47%. This rate was much lower than the anticipated response rate based on historic survey data. However, the previous re-interview experiment was done in 2011, so that expectations were very uncertain. As a consequence, there is room for selective nonresponse

between first and second interview. We start by looking at bivariate associations between re-interview response and a range of variables including both administrative variables and Health Survey variables. We then move to a logistic regression model explaining re-interview response.

In Table 5, we give the Cramer's V for auxiliary variables and Health Survey variables against the binary indicator of F2F re-interview response. The survey variables are measured in the first web interview. We consider response relative to the first interview response, so that survey variables can be treated as covariates. See Appendix B for a description of the code books of the auxiliary variables.

We fitted a logistic regression model including all variables from Table 5. We used a forward-backward selection procedure based on AIC to in/exclude covariates. The resulting model has an Nagelkerke R2 of 3.82% and includes the following variables:

Response F2F ~ Migration background + Age + Type of household.

We conclude that the response to the F2F re-interview was selective, but that associations with survey variables are weak and vanish once auxiliary variables are included. While there is no full guarantee, these results point at the potential to neutralize any impact of the first interview on response.

The web re-interview had a slightly higher response rate of 52%. Table 5 also includes the Cramer's V for the web re-interview. The fitted logistic regression model has a Nagelkerke R2 of 12.4% and has the following dependent variables:

Response WEB ~ Migration background + Age + Type of household + Generation + Gender.

The results are in line with those for the F2F re-interview. In fact, given that we find similar patterns, we feel safer in concluding that the mode itself did not have a strong impact on re-interview participation. We, thus, see support that we can use the web re-interview to evaluate structural time effects in health statistics.

As a final step, we look at the estimated time effects on the health statistics, following the strategy of Section 3.2. Table 6 shows the estimated time effects for the Health Survey variables. A few of the statistics show a significant time effect. However, when adjusting for multiple testing, these significances vanish. Based on this, we will not adjust estimated mode-specific measurement biases for time effects (as described in Section 3.2) in order to avoid inflation of the variance of estimated measurement biases.

Table 6: Estimated time effects $y_{web,web}^{R2cal} - y_{web,web}^{R1}$ on health survey statistics based on the web re-interview. Standard errors are estimated using bootstrap resampling.

Variable/statistic	Estimated time effect	Estimated standard error
Contact dentist (last 12 months)	0.0%	0.9%
Contact GP (last 12 months)	-0.1%	1.7%
Contact physio/exercise therapist (last 12 months)	-1.4%	1.2%
Diabetes type 2	0.4%	0.3%
Weekly sporter	-2.8%	1.3%
Sufficient fruit during 7 days a week	-0.1%	1.4%
Sufficient vegetables during 7 days a week	-3.6%	1.4%
Sufficiently active at moderate intensity	-4.6%	1.9%
Sufficient balance exercises	0.9%	1.5%
Sufficient muscle-bone strengthening activities	1.7%	1.6%
Use of non-prescribed medicine(s) (last 14 days)	-0.1%	1.8%
Smokers	-1.0%	0.7%
Severe overweight	-0.6%	0.8%
Heavy drinkers	0.0%	0.9%
Use of cannabis (last year)	-1.3%	0.7%
Use of prescribed medicine(s) (last 14 days)	-0.8%	1.4%
Self-perceived health (very) good	-0.1%	1.3%
Psychological distress (MHI-5<60)	0.6%	1.3%
Persons with ≥ 1 OECD limitation	1.5%	0.9%

4.2 What proportion of mode-specific biases can be attributed to measurement?

We start with the general research question on mode-specific measurement bias. For the four statistics that were selected to create re-interview strata (see Tables 2 and 3), Smoking – Non-prescribed medicine – Obese – Mental health, we explain in detail how the estimates are produced. In Section 4.3, we expand to the broad range of Health Survey statistics.

Table 7 displays the response means for the web first interview and for the F2F re-interview. The F2F re-interview response is calibrated to the web first interview response. In Section 4.2, we concluded that the strongest F2F re-interview response predictors are Migration background, Age, and Type of household. Our adjustment strategy is to add the Health Survey variable measured at first Web interview and apply a Generalized Regression Estimator (GREG) to adjust for re-interview nonresponse. So the weighting models become:

Smoker: Migration background + Age + Type of household + Smoker.Web.

Non-prescribed medicine: Migration background + Age + Type of household + Medicine.Web.

Overweight: Migration background + Age + Type of household + Overweight.Web.

Psychological distress: Migration background + Age + Type of household + Distress.Web.

Table 7 includes the adjusted estimates after applying GREG weighting. The mode-specific measurement bias estimate is the adjusted F2F response mean minus the web first response mean. The last column shows the resulting estimates.

Table 7: Response means web first interview and F2F re-interview, calibrated F2F re-interview means and estimated mode-specific measurement biases for Smoking, Non-prescribed medicine use, Obese and Mental health. Standard errors are based on bootstrap resampling.

Statistic	Response mean			ME bias estimate (SE)
	Web first (SE)	F2F unadjusted (SE)	F2F adjusted (SE)	
Smokers	11.8% (1.0%)	9.3% (1.3%)	11.2% (1.3%)	- 0.6% (1.0%)
N-p medicine	40.7% (1.6%)	56.8% (2.2%)	55.9% (2.1%)	15.2% (2.2%)
Overweight	11.4% (1.0%)	8.6% (1.3%)	9.4% (1.1%)	- 1.8% (0.9%)
Psych distress	13.4% (1.1%)	6.7% (1.2%)	8.0% (1.3%)	- 5.4% (1.3%)

We can now answer the first research question by concluding that mode-specific measurement biases are found but vary relatively a lot between survey variables.

In Section 4.3, we try to get a better understanding of this variation in biases by looking at the type of questions/statistics.

4.3 Are the directions of mode-specific measurement biases in line with expectations?

In this section, we consider our other two research questions, linking the directions of mode-specific measurement biases to sensitivity and complexity of questions as presented in Table 1.

We apply the estimation strategy elaborated in Section 3.1 and demonstrated in Section 4.2 to the full set of survey variables. Table 8 displays the results. The third column presents the estimated measurement difference relative to F2F for the population of respondents. The fourth column is the corresponding standard error. The fifth column is the actual change in statistics in case the measurement difference is corrected in the population of web plus F2F response. For example,

we find a difference of 15.2% of web relative to F2F in the use of non-prescribed medicine. If we subtract the difference and mix the resulting web estimate with the F2Festimate (that is not adjusted as it is the benchmark), then we get a12.1% change in the statistic.

Table 8: Estimated mode-specific measurement biases on health survey statistics based on the F2F re-interview. Standard errors are estimated using bootstrap resampling. Also given is the measurement bias adjustment on the full mixed-mode design taking F2F as measurement benchmark.

Type	Statistic/question	ME bias		MM adjustment
		Estimate	SE	
Robust	Contact dentist (last 12 months)	0.6%	1.1%	+0.4%
	Contact GP (last 12 months)	2.0%	2.0%	+1.6%
	Contact physio/exercise therapist (last 12 m)	1.4%	1.5%	+1.1%
	Diabetes type 2	0.0%	0.3%	+0.0%
	Weekly sporter	0.7%	1.7%	+0.6%
Complex	Sufficient fruit during 7 days a week	7.2%	1.9%	+5.7%
	Sufficient vegetables during 7 days a week	2.2%	2.0%	+1.8%
	Sufficiently active at moderate intensity	-7.7%	2.2%	-6.1%
	Sufficient balance exercises	7.2%	1.5%	+5.7%
	Sufficient muscle-bone strengthening activities	0.5%	1.8%	+0.4%
	Use of non-prescribed medicine(s) (last 14 d)	15.2%	2.2%	+12.1%
Sensitive	Smokers	-0.6%	1.0%	-0.5%
	Severe overweight	-1.8%	0.9%	-1.5%
	Heavy drinkers	-0.5%	1.2%	-0.4%
	Use of cannabis (last year)	-0.5%	0.6%	-0.4%
	Use of prescribed medicine(s) (last 14 days)	3.1%	1.7%	+2.5%
Subjective	Self-perceived health (very) good	5.8%	1.4%	+4.6%
	Psychological distress (MHI-5<60)	-5.4%	1.3%	-4.3%
	Persons with ≥ 1 OECD limitation	-1.2%	1.0%	-1.0%

Although standard errors are fairly large, clear patterns emerge from Table 8. For all the robust statistics the bias estimates are small. The complex statistics tend to show much larger bias estimates. In particular, the bias estimate for non-prescribed medicine is very large. Consultation of interviewers revealed that indeed this question often elicits clarification questions from respondents. The estimates of the complex statistics point in different directions, however. When interviewed in-person fruit consumption and the proportions satisfying the

balance fit nors go up. The broader, general fit norm gets worse and more non-prescribed medicine is used. We speculate that interviewers have an impact on these statistics, either by clarification or by keeping respondents more concentrated. The sensitive statistics show relatively small bias estimates, but we reiterate that two of them (heavy drinker and ever used drugs) are computer-assisted-self-interviewing for F2F. The subjective questions again show large bias estimates. These are in line with social desirability. In F2F, a more positive picture emerges, i.e. fewer issues and barriers and better self-perceived health. The biases of these statistics seem to have an opposite sign of the selection bias tendency when going from web to F2F. F2F attracts less healthy respondents but respondents are also more positive.

So what would it mean for Health Survey statistics if we were to adjust for the mode-specific measurement bias taking F2F as benchmark? In the Health Survey, approximately one quarter of response comes from F2F and three quarter comes from web. An adjustment towards F2F as measurement benchmark, consequently, implies bigger shifts than adjustment towards web. The last column of Table 8 has the bias-adjusted estimates when using F2F as benchmark. We see that the overall picture of health and healthy lifestyle gets more negative. The robust statistics already displayed this, but now also the other measurement bias adjusted statistics show a downward trend. Furthermore, and perhaps more importantly, as a result, the health statistics become more consistent. This will especially be crucial when comparing health statistics of subpopulations that have varying mode distributions, e.g. younger versus older.

5. Discussion

With the re-interview study we set out to answer three research questions that are important for the future of the Dutch Health Survey and other related surveys: 1) What proportion of mode-specific biases can be attributed to measurement?, 2) Are the directions of mode-specific measurement biases in line with social desirability?, and 3) Do the mode-specific measurement biases point at a role for interviewers in complex questions? These questions are the stepping stone to a review of the questionnaire design and to a reconsideration on the adaptive mixed-mode strategy. The important question how biases differ across relevant population strata was not discussed and is left for a follow-up study.

The re-interview experiment points at sizeable mode-specific measurement biases in health surveys. The prevalence of biases in statistics is smaller or larger depending on classifications of corresponding survey questions based on subjectivity, complexity and sensitivity. In general, these variations are in line with expectations. Statistics deemed to be robust to mode indeed showed small biases and complex and subjective questions showed larger biases. The biases found for subjective questions correspond to social desirability, i.e. they are more positive when asked in-person. This does not necessarily mean that there is social

desirability bias. However, the contradiction in health statistics, where F2F responses point at a less healthy life style and more use of health care but no difference in self-perception, seems paradoxical. The questions categorized as complex display biases in different directions. For at least one of the questions, non-prescribed medicine, we know that interviewers report that respondents are uncertain about the exact definition. The other 'complex' statistics on physical activity and nutrition are based on series of questions and we speculate that interviewers either assist respondents or keep them motivated.

We must stress that the re-interview study relies on various assumptions that can only be checked in part. We added a control group in order to explore time change and experimental impact, but found little evidence for such change or impact. Furthermore, the key Health Survey variables turned out to be (very) weakly related to participation in the re-interview. This is an important finding as such associations would have meant we need to rely more strongly on the effective calibration of re-interview response. Nonetheless, it must be accepted that assumptions are violated at least to some extent. If so, estimates for relative measurement biases may be larger than they really are.

There are a few other limitations of our study. The statistical power was set at observable biases of 5%. This is fairly large and subtle bias patterns cannot be detected. It implies that there may be inconsistencies in biases for questions that have similar characteristics. Also, because of budget restrictions, we limited the re-interview to the strata that were at risk most, i.e. showed the largest differences between modes. These made up around half of the population. This makes it hard to extrapolate to the other half of the population. Finally, this study was set in the Dutch context. While we speculate that the measurement classification of questions is relatively stable across countries, modes can be implemented differently across countries. Also mode coverage and mode preferences differ from one country to the other.

This study marks the beginning of a series of follow-up activities. First, given the relation between question characteristics and the size of biases, it is imperative that also other surveys are evaluated; in particular, those that contain similar types of questions. Before making a decision to go for a re-interview experimental design, first conjectures about biases may be formulated through potential inconsistencies in associations between statistics and classifications of questions. To bridge to other survey settings, ideally such evaluations are done in multiple countries. Directly related to this extension is a review of bias size and directions with questionnaire experts and designers and with face-to-face interviewers. Questionnaire designers and interviewers may confirm that certain questions require more cognitive effort or are subject to some confusion about definitions and scope. At the time of writing, the survey question on non-prescribed medication had just been rephrased and implemented with the intention of reducing mode differences. From there, an obvious next step is to bias estimates for a selection of population subgroups. Also, estimates for mode-specific measurement biases should be made taking web as the measurement benchmark. Preliminary estimates for the web benchmark can be made fairly quickly using the mode distributions. However, given that the biases likely are heterogeneous about

persons with different backgrounds, the actual estimates will show subtle rearrangements relative to the preliminary estimates. Finally, a number of adaptive survey design scenarios need to be considered. We see three scenarios: One scenario is that within the adaptive survey design the shares of web and face-to-face response are stabilized in time. This is a cheap and on the longer run not sustainable solution. Another scenario is that biases are adjusted and the utility of going adaptive per stratum is revisited. A final scenario is that the adaptation itself explicitly accounts for the measurement biases. In a separate paper, we will elaborate the scenarios.

References

Berkel, K. van, Doef, S. van der, Schouten, B. (2020), Implementation of adaptive survey design with an application to the Dutch Health survey, Journal of Official Statistics, 36 (3), 609 – 629.

Beukenhorst, D.J., Buelens, B., Engelen, F., Laan, J. van der, Meertens, V. and Schouten, B. (2013), The impact of Survey item characteristics on mode-specific measurement bias in the Crime Victimisation Survey, Discussion paper 201416. Statistics Netherlands, Den Haag.

Calinescu, M. (2013), Optimal Resource Allocation in Adaptive Survey Designs, PhD thesis, VU University Amsterdam, Faculty of Exact Sciences, Department of Mathematics and Computer Science, Amsterdam, The Netherlands, available at https://research.vu.nl/publications/optimal-resource-allocation-in-adaptive-survey-designs .

Campanelli, P., Nicolaas, G., Jäckle, A., Lynn, P., Hope, S., Blake, M., and Gray, M. (2011), A classification of question characteristics relevant to measurement (error) and consequently important for mixed mode questionnaire design, Paper presented at the Royal Statistical Society, October 11, London, UK.

Dillman, D.A., Smyth, J.D., and Christian, L.M. (2014), Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailor Design Method, 4th edition, New York: Wiley and Sons.

Klausch, L.T., Schouten, B., Buelens, B., Van den Brakel, J. (2017), Adjusting measurement bias in sequential mixed-mode surveys using re-interview data, Journal of Survey Statistics and Methodology, 5 (4), 409 – 432.

Kreuter, F., Presser, S. and Tourangeau, R. (2008), Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity, Public Opinion Quarterly, 72 (5), 847-865.

Lensvelt-Mulders, G. J. L. M. (2008), Surveying sensitive topics, Pages 461-478 in International Handbook Of Survey Methodology, edited by E. D. de Leeuw, J. J. Hox, and D. A. Dillman. New York: Taylor & Francis, Psychology Press, EAM series.

Roberts, C., Vandenplas, C. (2017), Estimating Components of Mean Squared Error to Evaluate the Benefits of Mixing Data Collection Modes, Journal of Official Statistics, 33 (2), 303–334.

Sakshaug, J., Cernat, A., Raghunathan, T.E. (2019), Do Mixed-Mode Surveys Decrease Nonresponse Bias, Measurement Error Bias and Total Bias. An Experimental Study, Journal of Survey Statistics and Methodology, 7, 545–571.

Saris, W. E. and Gallhofer, I. (2007), Estimation of the effects of measurement characteristics on the quality of survey questions, Survey Research Methods, 1(1), 29-43.

Schouten, B., Brakel, J. van den, Buelens, B., Giesen, D., Luiten, A., Meertens, V. (2021), Mixed-Mode Official Surveys: Design and Analysis, Chapmann & Hall/CRC.

Schouten, B., Klausch, L.T., Buelens, B., Brakel, J van den (2024), A cost-benefit analysis of re-interview designs for estimating and adjusting mode measurement effects, Journal of Survey Statistics and Methodology,12 (3), 790–813.

Schouten, B., Peytchev, A., Wagner, J. (2017), Adaptive Survey Design, Series on Statistics Handbooks, Chapman and Hall/CRC.

Tourangeau, R. and Yan, T. (2007), Sensitive Questions in Surveys, Psychological Bulletin, 133 (5), 859-883.

Vaart, W. van der, Zouwen, J. van der, and Dijkstra, W. (1995), Retrospective questions: data quality, task difficulty, and the use of a checklist, Quality and Quantity, 29 (3), 299-315.

Yu, W., Elliott, M.R, Raghunathan, T.E. (2024), Three Approaches to Improve Inferences Based on Survey Data Collected with Mixed-Mode Designs, Journal of Survey Statistics and Methodology, 12(3), 814–839.

Zouwen, J. van der (2000), An assessment of the difficulty of questions used in the ISSP-questionnaires, the clarity of their wording and the comparability of the responses, ZA-Informationen, 45, 96-114.

Appendix A - Interviewer training

Here, we list a number of questions that were anticipated respondents would ask at the door when performing the re-interview. To each question an answer was suggested.

Q: Last time, I filled out the survey online. Can I do that again now?
A: No, the survey protocol requires the second questionnaire to be conducted face-to-face.

Q: Why is it with an interviewer this time?
A: This is a one-time occurrence. It's a smaller study to which we are dedicating more attention.

Q: Nothing has changed for me since the last time, so I'm no longer interesting, right?
A: Quite the opposite. If only people participate whose situations have changed, we would obtain an inaccurate representation.

Q: Can you (the interviewer) see my answers from last time?
A: No, due to privacy concerns, answers from the previous time are not disclosed.

Q: I can't remember participating last time. Is that correct?
A: According to our registration, it's been six to eight weeks since you participated.

Q: I recently filled out a questionnaire, why do I have to participate in another survey?
A: For some, a lot can change in a few weeks, while for others, nothing changes. To gain more insight into the health of the residents of the Netherlands, we are asking a small number of people to participate once more in the survey.

Appendix B - Auxiliary variables linked from administrative data

Table B.1 contains the auxiliary variables in logistic regression models and re-interview weighting adjustment.

Table B.1: List of auxiliary variables and code book.

Variable	Categories
Type of household	Single household, Couple without children / Other household, Couple with children / with Unknown, Single-parent household.
Place in household	Child living at home, Single, Partner without children / Reference person in other household, Other member of a household / Unknown, Partner without children / Parent in a single-parent household.
Socio-economic status	Employee, Director-major shareholder / Self-employed entrepreneur / Other self-employed person / Family worker, Recipient of unemployment benefit / Benefit recipient / Recipient of social benefit other / Recipient of sickness invalidity benefit, Recipient of pension allowance, Under-school or school student with income, Under-school or school student without income, Other without income / Belongs to household with no observed income / Unknown.
Age	0–11, 12–24, 25–44, 45–64, 65 +
Province +	Groningen, Friesland, Drenthe, Overijssel, Flevoland, Gelderland, Utrecht excluding municipality of Utrecht, Noord-Holland excluding municipality of Amsterdam, Zuid-Holland excluding municipalities of Den Haag and Rotterdam, Zeeland, Noord-Brabant, Limburg, Municipality of Utrecht,

	Municipality of Amsterdam, Municipality of Den Haag, Municipality of Rotterdam.
Migration background	Dutch, Non-Western migrants, Western migrants.
Type of owner	Owner lives in the house, The landlord is a housing association, Landlord other than housing association.
Marital status	Married, including registered partnership, Never been married, Divorced, Widow / widower.
Province	Groningen, Friesland, Drenthe, Overijssel, Flevoland, Gelderland, Utrecht, Noord-Holland, Zuid-Holland, Zeeland, Noord-Brabant, Limburg.
Household income	1−20 percentile, 21−40 percentile, 41−60 percentile, 61−80 percentile, 81−100 percentile.
Urbanicity	Very strongly urban, Strongly urban, Moderately urban, Little urban, Not urban.
Welfare	1−20 percentile, 21−40 percentile, 41−60 percentile, 61−80 percentile, 81−100 percentile.
Number in household	1, 2, 3, 4, 5 or more.
Part of the country	North-Netherlands, East-Netherlands, South-Netherlands,

	West-Netherlands.
Gender	Male, Female.
Generation	Dutch, First generation migrant, Second generation migrant.

Colophon

Publisher
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
Statistics Netherlands, CCN Creation and visualisation

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbs.nl/information