Assessing the mode-dependency of sample selectivity across the survey response process

Thomas Klausch, Joop Hox and Barry Schouten

Discussion paper (201303)

Statistics Netherlands

The Hague/Heerlen, 2013

Explanation of symbols

.	data not available
*****	provisional figure
******	revised provisional figure (but not definite)
x	publication prohibited (confidential figure)
–	nil
–	(between two figures) inclusive
0 (0.0)	less than half of unit concerned
empty cell	not applicable
2012–2013	2012 to 2013 inclusive
2012/2013	average for 2012 up to and including 2013
2012/’13	crop year, financial year, school year etc. beginning in 2012 and ending in 2013
2010/’11–2012/’13	crop year, financial year, etc. 2010/’11 to 2012/’13 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

60083201303 X-10

Assessing the mode-dependency of sample selectivity across the survey response process

Thomas Klausch, Joop Hox and Barry Schouten

Summary: Survey researchers currently act on three beliefs when combining survey modes in mixed-mode designs. First, modes elicit distinct patterns of survey response behaviour and nonresponse bias. Second, these selection differences are caused by differences in the response process of differential coverage, contact and cooperation. Third, mode-dependent response patterns might be exploitable by sequential mixed-mode designs ideally yielding samples less biased by selection and hence more 'representative'. These assumptions are assessed using a factorial design, in which the Dutch Crime Victimisation Survey was administered either in CAPI, CATI, mail, or web. By means of a second wave we gained insight into wave 1 web coverage, contact, bias on target variables and sequential mixed-mode (non)response. Results differ by measure of selectivity, assessed either by response rates, selection bias, or representativeness indicators. Differences in response rates were not reflected by patterns of relative sample bias or representativeness, which were generally small. Selection was homogenous for the two interviewer modes and self-administered modes, respectively, but a bit more heterogonous between interviewer and self-administered modes. Mode-dependent selection bias on target variables was not identified. Representativeness analyses showed that CAPI and, surprisingly, web, were most representative. This was found despite strong impact of non-coverage on web representativeness, which was mitigated in the process of contact and cooperation for web. Sequential designs increased response rates. They did not produce more, but also not less representative samples.

Keywords: Mode effects; Nonresponse; Nonresponse bias.

1. Introduction

Modern survey research targeting national or international general populations is currently facing serious challenges in the nonresponse arena. Declining response rates have caused worries about increased imprecision of estimates due to nonresponse bias (De Heer, 1999; De Leeuw & De Heer, 2002). An underlying cause to this trend is increasing refusal to survey modes that, traditionally, reliably yielded high response rates, such as face-to-face or telephone surveys (Hox & De Leeuw, 1994). Another part of the problem is represented by decreasing telephone coverage of the general public (Blumberg, Luke, Cynamon, & Frankel, 2008; Steeh, 2008), while household equipment with internet access is still incomplete in many countries (Blyth, 2008; De Leeuw & Hox, 2008, p. 250-251; Manfreda & Vehovar, 2008, p. 269; Mohorko, 2011). Unimode surveys using telephone or the web as communication channel during data collection are, therefore, being considered insufficient to cover the whole sampling frame (e.g. Blyth, 2008), while also face-to-face or mail surveys can now be suspected to yield biased samples. Undercoverage is particularly problematic, if surveys serve as basis for democratic policy making. In this case non-coverage suppresses information from a share of the population, threatening their legitimacy as democratic tool. Finally, cost considerations are another problem, urging practitioners to reduce gross samples of the cost intensive face-to-face mode, for example.

Survey researchers all over the world have started to use mixed-mode survey designs in an effort to address these deficits. For example, survey modes are ordered in sequence presenting nonrespondents of one survey mode with an alternative mode as response channel (De Leeuw, 2005; Dillman & Christian, 2005; De Leeuw, Dillman, & Hox, 2008; Dillman, Smyth, & Christian, 2009, p. 300-310; De Leeuw & Hox, 2010). The choice to mix survey modes is eventually motivated by the hope that more than one mode combined in a survey project would compensate the deficits of any single, possibly cheaper, survey mode in the joint net sample. That is, mixed-mode surveys might produce net samples, which are at least as 'representative' or more by reflecting the target population with equal or less bias than any single-mode sample would (Dillman, Smyth, & Christian, 2009, p. 303-305).

Ordering of survey modes in sequence is often found to be successful in yielding quite substantial net increases in response (Link & Mokdad, 2006; Converse, Wolfe, Xiaoting, & Oswald, 2008; Dillman, Phelps, et al., 2009). This finding by itself is, however, not truly surprising, since it is well-known that the expected response rates of surveys differ by mode. Hox and De Leeuw (1994) showed already some twenty years ago that face-to-face surveys generally produce higher response rates than mail and telephone surveys do. More recent meta-analyses report that web surveys generally do worse than other survey modes (Shih & Fan, 2008; Manfreda, Bosnjak, Berzelak, Haas, & Vehovar, 2008). It is useful to note that such effects might be simply constant across the target population, which by itself may explain the net increase of response rates in mixed-mode surveys. Moreover, if this was so, analysts could take the effects a priori into account when designing single-mode surveys, e.g. regarding desired power of statistical tests, and there would not be an immediate need for a mixed-mode design.

Mixed-mode, therefore, requires more to achieve the objective set out above. The crucial main assumption of mixed-mode surveys, often made implicitly, is that different people reply in different modes – survey (non)response is mode-dependent. Without mode-dependency of response the utility of mixed-mode surveys is strongly reduced, because surveys using any single mode would yield similar samples, while marginal response probabilities might simply differ by a fixed effect. Since mixed-mode surveys additionally cause a lot of effort in administration and also bring about new problems, for example concerning the handling of mode-dependent measurement error (De Leeuw & Hox, 2010; Jäckle, Roberts, & Lynn, 2010; Vannieuwenhuyze, Loosveldt, & Molenberghs, 2010; Heerwegh & Loosveldt, 2011), it is imperative to assure that survey response is indeed mode-dependent instead of mode-independent. The property is often assumed by survey practitioners using mixed-mode designs trying to maximize survey response rates. However, response rates alone provide insufficient certainty for the absence of sample bias (Groves, 2006; Groves & Peytcheva, 2008). Therefore, we intend to provide a more detailed assessment in the present study.

Why should survey-response be mode-dependent after all? Many researchers would argue that this mainly relates to differences in the response process of coverage, contact and cooperation (e.g. Groves, 1989). Possibly some researchers put more emphasis on coverage differences, others on cooperation or just consider overall differences. In this study, we intend to dissect the process in more detail to show, where mode-dependent differences are, and if common beliefs on the mode-dependency of response and nonresponse biases are confirmed. We question whether survey response really is mode-dependent, yielding mode-dependent nonresponse bias. We ask, which role differential nonresponse types, coverage, contact and refusal, take on in the response process, and if common sequential mixed-mode designs can exploit mode-dependence of response.

To answer these questions, we present empirical results from a randomized factorial experiment conducted at Statistics Netherlands in 2011. Over the last years Statistics Netherlands has redesigned many of its social surveys from single mode designs to multi-mode designs that include web. We refer to Cuppen, Van der Laan and Van Nunspeet (2011) and Schouten (2012) for details about the redesign projects. The experiment was conducted within a larger research project about mode effects in social surveys. Buelens et al (2012) present detailed descriptions of mode effect decompositions for a number of key survey variables from the Labour Force Survey and Crime Vicitmisation Survey. Here, we restrict ourselves to mode-specific nonresponse.

While we are not the first to use experiments to compare modes, many past studies could provide less detailed conclusions as our data allow (section 3). Throughout this paper we concentrate, theoretically (sections 2 to 4) and empirically (sections 5 and 6), on the four major survey modes: face-to-face, telephone, mail, and web interviews. We acknowledge that this is a rather wide classifications and there are many possible differences in specific designs.

It should also be acknowledged that mode differences in the measurement process have been discussed as another important source of bias contributing to net differences between modes (e.g. Jäckle, Roberts, & Lynn, 2010). We believe that measurement processes are equally important as mode-dependent nonresponse in the context of mixed-mode surveys, but the scope of this paper is limited to an analysis of nonresponse. Moreover, we either use register data or data collected in a single survey mode for the assessment of mode-dependent nonresponse. Our validation data, therefore, allow an assessment of nonresponse, which is independent of a mode-specific measurement process (register data) or limited to a particular measurement mode (single-mode data).

2. Analytical description of mode-dependent survey response and nonresponse bias

Clear definitions of mode-dependent nonresponse are generally absent in the literature, despite the large body of research comparing survey modes (cf. section 3). Recently, Vannieuwenhuyze, Loosveldt, and Molenberghs (2010) suggested definitions, which apply in a comparative situation of sequential and single mode survey designs. We intend to provide a suggestion for a non-comparative situation. We suggest a definition along the lines of traditional nonresponse research (e.g. Särndal & Lundström, 2005) reviewing the regular single-mode approach to nonresponse bias, before extending it to a multimode situation.

2.1 The single-mode description of survey response and nonresponse bias

We assume a random response model, which suggests that each population element *i* has a fixed probability to be in the response sample (e.g. Bethlehem, 1988, 2002), $P(R_i = 1) = \rho_i$ where R_i is a Bernoulli random variable taking the value 1, if population element *i* responds if it is sampled. For a given target variable *y*, the bias in its mean due to nonresponse is then given by

$$B(\bar{y}) = \frac{\sigma_{y,\rho}}{\bar{\rho}} \leq \frac{\sigma_y \sigma_\rho}{\bar{\rho}} \tag{1}$$

where $\sigma_{y,\rho}$ denotes the population covariance between y and ρ and $\bar{\rho}$ its mean. $\bar{\rho}$ also represents the expected response rate. Bounds to the bias are given by the inequality, in which σ denotes the standard deviation of *y* and ρ respectively (Schouten et al., 2009). From (1) it can be seen that bias is absent, if the variance of response probabilities is zero, i.e. if response probabilities are constant in the population, for a given survey with a given design, say. Increased variability increases expected bias, which has led Schouten and colleagues to formulate a representativeness (R-) indicator based on the variability of response probabilities, defined as

$$R(\rho) = 1 - 2\sigma_\rho \tag{2}$$

In practice, the R-indicator is a summary measure for a number of auxiliary variables, on which bias is assessed and representativeness should be summarised. It is a model-based estimate. Observing that the covariance between a survey's target variable *y* and the response probability is determinant of bias, Groves (2006) and Groves and Peytcheva (2008) suggested the causal response model. Underlying covariates causally influence the response probabilities and target variables. Let *X* be a set of possible causes to the response probability of element *i*, then

$$P(R_i = 1 | X) = link\,(X_i \beta')^{-1} \tag{3}$$

Any sample realized by process (3) will be biased for any *X* related to ρ_i, following (1). This includes the survey target variables, *X*=*y*, of course. If $\beta \neq 0$ for any *X*, and *y* is independent of *X*, no bias on *y* is caused. On the opposite, if $\beta = 0$ for any *X*, there still may be bias on *y*, if *y* itself is causally related to the response indicator. A special case is represented by bias caused by a spurious relation between *R* and *y* induced by *X* (Groves, 2006).

Furthermore, in appreciating the fact that survey response is a stepwise process (cf. section 4), we can define probabilities for coverage, contact, and cooperation. In doing so, we assume that being covered is in fact a probabilistic event, which conforms to assumptions of the random response model. Let C_1, C_2, C_3 be random binary variables indicating success if coverage, contact, and cooperation are given for individual *i*. Then we have

$$P(R_i \mid X) = P(C_{i1} \cap C_{i2} \cap C_{i3} \mid X) = P(C_{i1} \mid X)P(C_{i2} \mid X, C_{i1})P(C_{i3} \mid X, C_{i1}, C_{i2}) \quad (4)$$

which can be modelled either jointly or separately to explore where bias is created.

2.2 Defining mode-dependent survey response and nonresponse bias

Extensions to mode-dependent selectivity and bias are straightforward. However, now we assume that each population element has a set of *M* distinct probabilities to respond in each of *M* survey modes. Moreover, we assume that design properties, such as mode of first contact, survey topic and fieldwork organization, are constant across modes. For the mode-dependent response probability to mode *M=m* we have:

$$P_m(R_{im} \mid X) = P_m(C_{im1} \mid X)P_m(C_{im2} \mid X, C_{im1})P_m(C_{im3} \mid X, C_{im1}, C_{im2}) = \rho_{im} \quad (5)$$

Obviously, the response probabilities ρ_{im} are the result of a complex interaction of causes on each of the three stages. We can describe mode-dependent nonresponse bias by

$$B(\bar{y}_{m=1}) - B(\bar{y}_{m=2}) = \frac{\sigma_{y,\rho_1}}{\bar{\rho}_{m=1}} - \frac{\sigma_{y,\rho_{i2}}}{\bar{\rho}_{m=2}} \quad (6)$$

Based on (6) we can derive 3 definitions for mode-dependent response. Literature has often used the response rate to compare survey response (cf. section 3). That is why the first definition is based on expected response rates:

Definition 1 (Mode-dependent expected response rates): A given design produces a mode-dependent survey response for two survey modes, if mean response probabilities differ, i.e.

$$\bar{\rho}_{m=1} = \bar{\rho}_{m=2}$$

However, definition 1 is only meaningful in the assessment of bias if there is a relationship between a variable of interest and the response propensity distribution. Therefore, we define response to be mode-dependent conditional on cause *X* at stage *C* with respect to two given modes.

Definition 2 (Mode-dependent relative selection bias): A given design produces a mode-dependent response process and relative selection bias for variable X, if associated parameters in a regression model of ρ_{im} *are vary over two survey modes, i.e.*

$$\beta_{C,m=1} \neq \beta_{C,m=2}$$

Note that definition 2 directly implies selection bias, following (6) on the associated *X*. Since even small effect sizes would still have us conclude that mode-dependency in response is present, this definition is rather strict. Moreover, in practice statistical power will influence the analyst' ability to identify definition 2. Hence, careful consideration of effect sizes qualitatively and quantitatively remains to the analyst. It is clear from this definition that mode-dependency may be formulated conditional on a particular nonresponse type *C* or on the overall outcome *R*.

We can also take from (1) that mode-dependent bias is absent, if either both modes yield fully representative samples, indicated by constant variance of both mode-dependent response probabilities, or if both modes yield equally biased samples with respect to *Y*, indicated by equal covariance between *Y* and the response propensity. Because of the strictness of the first definition, we use this result to offer a weaker definition for mode-dependent response based on the R-indicator. We suggest that two modes yield equally representative samples if the variability in response probabilities in the population is equal.

Definition 3 (Mode-representativeness): A given design produces a mode-representative response process, if the variation in propensities over a set of auxiliary variables X is equivalent for two survey modes, i.e.

$$S(\rho_{m=1}) = S(\rho_{m=2})$$

Definition 3 can be assessed using R-indicators. Every definition suggested is based on the analytic description of nonresponse bias. Definition 1 has the least implication for bias, whereas definitions 2 and 3 are directly related. However, definitions 2 and 3 need not lead to the same conclusions in practice. Definition 2 is requires statistical tests for each considered *X* (cf. section 5). Definition 3, based on the R-indicator, uses a summary statistic that weights deviations from mean response probability more strongly. Definitions 1 and 3 combine to (mode-dependent) maximal bias, not studied in detail here (cf. Schouten et al., 2009).

3. Evidence for the mode-dependency of survey response and nonresponse bias

The general research question of our study can now be stated as follows.

> RQ1: *In which way do expected response rates, relative sample bias and representativeness depend on the mode of data administration?*

Since response rates today are established as quality indicators of surveys, survey modes with lower response rates are suspected to be biased more strongly and less representative. Three available meta-analyses suggest a general hierarchy of response rates: CAPI, CATI, mail, and web, in decreasing order of response, say (Hox & de Leeuw, 1994; Shih & Fan, 2008;

Manfreda, Bosnjak, Berzelak, Haas, & Vehovar, 2008). Thus, one would expect web surveys to yield lowest response rates, for example, and, hence, strongest bias or least quality. However, as Groves and Peytcheva (2008) showed, response rates and nonresponse bias are at best weakly related. Therefore, such a conclusion seems at least doubtful.

Sample bias, on the other hand, can theoretically be assessed on many types of variables, provided their availability in the sample and the population. Mode comparison studies have compared the samples realized by different modes of administration. However, reviewing past studies in this field it is very hard to synthesize generalisable findings. In the past decade, there has been great interest in comparing web with either telephone or mail surveys. web samples are quite consistently found to be higher educated and more affluent than respondents in mail or telephone surveys (Miller et al., 2002; Bälter, Bälter, Fondell, Lagerros, 2005; Link & Mokdad, 2005, 2006; Dillman, Phelps, et al., 2009). However, regarding most other socio-demographics, such as age, gender, race, Marital Status, household sizes etc., results are very mixed and hard to synthesize in a coherent manner. Also, one may note that earlier findings are probably outdated, because of recent developments regarding internet and telephone coverage, changing population structures and survey-taking climates. Nevertheless, many studies do find mode-dependent response processes, though we cannot fully describe them consistently on a meta-analytic level.

We can think of some explanations for this inconsistency beyond sampling variation. First, some studies did not sample from the general population. Instead special interest groups were used as sampling frames, such as students or employees (e.g. Kwak & Radler, 2002; Sax, Gilmartin, & Bryant, 2003; Kaplowitz, Hadlock, & Levine, 2004). A second, related, problem may be differences in international population structures (e.g. US versus Europe). A third problem is pre-selection. Though many mode comparisons are conducted on US populations, some scholars had to use existing panel populations, recruited experimental units from a pre-study or pre-select samples in other ways (Fricker, Galesic, Tourangeau, & Yan, 2005; Dillman, Phelps, et al., 2009; Chang & Krosnick, 2009).

One can also differentiate relative from absolute assessments. Relative assessments compare modes against each other, whilst absolute assessments test response samples against a known population or sample marginal. The disadvantage about the relative perspective is unavailability of information on the extent of total bias in any mode. A second problem about the relative assessment is that it is often based on endogenous measurements obtained for each mode that is compared. Hence mode effects on measurement might be mistaken for differences in nonresponse.

All of the above limits studies' generalisability and comparability. The added value of our study comes from four key elements. First of all, to our knowledge research has never, as we write, focused all of the four most important survey modes in a single experiment (CAPI, CATI, mail, and web). Two-way mode comparisons are most common, but any prior studies that merely involve subsets of the major modes are very difficult to compare among each other. We will present the first overall assessment of the major modes. As we noted, we will additionally provide an assessment, of whether mixed-mode can exploit mode-dependent (non-) response. Moreover, we will apply a national probability sample, and present nonresponse analyses based on register data. Finally, we will not stop at an analysis of socio-demographics, but extend some of the conclusions to our survey's target variables.

Our experiment is limited in its scope to a particular survey topic, the mode-specific designs, and the survey taking climate in The Netherlands. We nevertheless hope to motivate that findings will apply to other situations and countries with similar populations in terms of structure and accessibility.

4. Sources of mode-dependent survey response

Survey response can be conceptualized as a sequence of conditions that a unit has to pass in order to be included in the response sample (Groves, 1989; Groves & Couper, 1998), while only on the last stage the (non)respondent is actively involved (Figure 1). Most importantly, units have to be covered by the mode of administration, successfully be contacted by a given contact mode (e.g. telephone or by mail) and cooperate in a given mode of administration[1] (e.g. face-to-face, by telephone, mail, or web). In prior research it has not been scrutinized, how the different modes contribute to overall differences in sample bias and representativeness. Therefore we ask:

> RQ2: *How is mode-dependent sample bias and representativeness created in the process of survey response (coverage, contact, cooperation)?*

4.1 Mode-dependency due to differential frame coverage

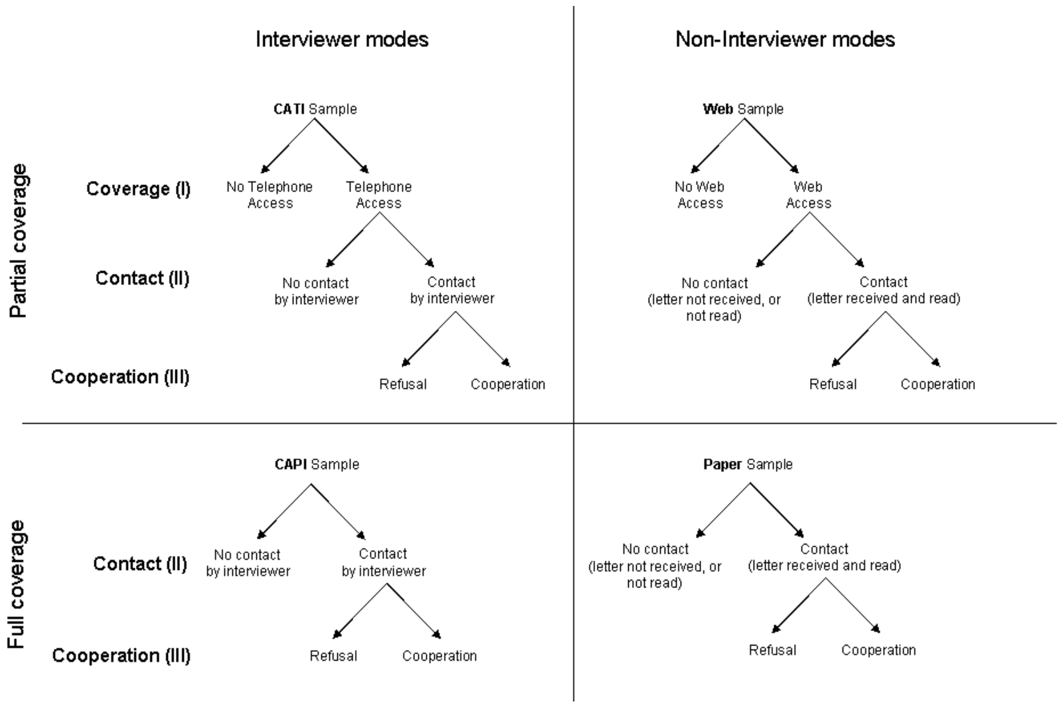
Coverage bias is caused by incomplete accessibility of population units by the sampling frame, while accessibility usually is strongly determined by the mode of administration. In telephone surveys the coverage problem (Figure 1, level 1) is a classical cause of selection bias (Groves, 1989, p.11; Biemer & Lyberg, 2003, p.64-68), which, however, more recently also received a lot of attention as chief limitation to web surveys of the general public (Couper, 2000). Dillman, Smyth, et al. (2009a, p. 46-49) also describe coverage problems in mail surveys using incomplete address lists. This observation, however, can be generalized to the problem when practitioners have to acquire different, only partially complete, lists or samples of contact data from different sources or providers. Assuming the existence of a complete 'list' of the sampling frame units, the coverage problem is limited to technical accessibility of persons and households. This includes cases when a population register is available. mail and face-to-face surveys can then hardly suffer from coverage bias, which is the situation assumed in figure 1 and henceforth.

If a sampled person cannot be reached in the absence of a landline phone, the factors causing and correlating with this state will cause coverage bias of the telephone mode vis-à-vis a mode with full population coverage, say CAPI or mail, or different population coverage, like

1. There are some other causes to mode-dependent nonresponse, in particular 'eligibility' and 'ability' (e.g. Lynn, 2008). For example, the 'reading' ability is a necessity for participation in self-administered surveys. In this paper, however, we focus on the three major reasons for mode-dependent survey nonresponse, i.e. non-coverage, non-contact and refusal. „Ability" to participate is not included in this graph, but inability represents another form of nonresponse. It can be seen as part of cooperation/refusal

the internet. In the US, for example, telephone coverage is higher for the non-poor, two or more person households in rural regions living in owned dwellings (Blumberg et al., 2008, p.82). These current data mainly coincide with trends already known from the 1980s for the US and Europe (Thornberry & Massey, 1988; Trewin & Lee, 1988). In Europe, household coverage with landline rates range from as low as 40% (Lithuania) to near full coverage (Sweden), according to estimates from the Eurobarometer 2005, a face-to-face survey (Blyth, 2008). Furthermore, authors found strong bias in landline-only groups in analyses based on the Eurobarometer regarding various socio-demographic and attitudinal indicators (e.g. Busse & Fuchs, 2009, 2011; Mohorko, 2011). They warned of the deleterious impact in surveys using only the landline telephone mode that neglect the 'mobile-only' population. This especially seems to hold, because bias tends to increase over time with decreasing coverage rates (Mohorko, 2011).

Figure 1: Illustration of mode-dependent survey response by nonresponse type



Surveys using the web as mode of administration have traditionally been considered useful only for specific sub-populations with high internet coverage (Manfreda & Vehovar, 2008). In fact, insufficient household coverage with internet access for a long time was considered the "biggest threat to the representativeness" of general population internet surveys (Couper, 2000, p. 467). It has been contemplated that 'the digital divide' was caused by quicker adoption of the internet by those population domains with higher socio-economic status, especially the higher educated, affluent, and young (Couper, 2000; Kwak & Radler, 2002; Berrens et al., 2003).

nonresponse, because non-able persons often refuse to participate. From a different perspective, non-able units do not belong to the sampling frame (are ineligible) and thus are outside of the graph pictured here.

However, contrary to landline telephone, household coverage with internet access is on the rise in most industrialized countries. It has therefore been described as a potentially vanishing problem (e.g. De Leeuw & Hox, 2008; Mohorko, 2011). For example in The Netherlands, our study population, the household coverage with internet access has reached about 94% (CBS, 2011; based on a national CATI survey on ICT usage). But in many other European countries web coverage is much lower.

Today coverage biases probably concern distinct sub-groups of the population. Couper, Kapteyn, Schonlau, and Winter (2007) showed, for example, that internet use in the US population aged 50 or above can still be explained well by a number of socio-demographics, especially higher income, lower age, higher education, as well as living in (sub)urban regions. web-coverage problem is becoming more and more a multi-dimensional issue, due to multiple browsers and new instruments, like mobile handhelds (e.g. Fuchs, 2008). In this study, we will limit web coverage to the availability of household internet access using a personal computer.

Expectations about the impact of coverage. We expect that the influence of non-coverage by landline telephone and internet access on mode-dependent bias and representativeness will be substantial. This is consistent with scholars' statements about the potentially deleterious effects of non-coverage on sample bias. For web and telephone it can be expected to find the typical socio-demographic patterns described above.

4.2 Mode-dependency due to differential processes in establishing contact

The mode of contact chosen may vary across possible design decisions and naturally may differ from the mode of administration. Dillman, Smyth, et al. (2009) advocate a series of 'tailored' contact steps ranging from pre-notification, cover letter to repeated reminders. However, ultimately an interviewer has to establish contact in the interviewer modes, whereas this step is absent in the self-administered modes. The success of contacting a person generally depends on the effort invested. In fact, a major strength of interviewer administered surveys is the additional effort made by a human to establish contact and remind the sampled person of his or her participation beyond sending invitation letters and reminders. Accordingly, non-contact rates are small in interviewer administered surveys, for example below ten percent in many European countries, which represents in some cases less than a twentieth of the nonresponse due to refusal (Billiet, Philippens, Fitzgerald, & Stoop, 2007, based on the ESS, a face-to-face survey). Time-dependent at home patterns and related characteristics, such as employment and leisure activities, influence the contact probability on the respondent side (Groves & Couper, 1998; Keeter, Miller, Kohut, Groves, & Presser, 2000; Lynn & Clarke, 2002; Nicoletti & Peracchi, 2005; Durrant & Steele, 2009).

Little is known about the selective processes due to non-contact in mail or web surveys, except for evidence on effects of different contact strategies by mail or eMail (e.g. Kaplowitz, Hardlock & Levin, 2004; Porter & Whitcomb, 2007). Generally, non-interviewer administered surveys are less obtrusive, because (pre-) notifications are more often overlooked, not read and forgotten. However, written notifications perhaps more often reach hard-to-contact persons, because letters can be received and read at times and situations inaccessible to interviewers. Hence conclusions drawn from interviewer administered studies need not hold for the non-interviewer modes. This may introduce mode-dependent bias on the contact stage.

Expectations about the impact of contact. There is little indication on the particular impact of non-contact on bias or representativity. Past research studying cooperation in interviewer surveys finds very similar correlates, irrespective whether CATI or CAPI surveys are considered. So we expect that bias due to noncontact will be similar in these modes. The interviewer is the outstanding difference between CATI/CAPI on the one hand and mail/web contact strategies on the other. We expect that this will lead to larger differences between these modes groups, respectively.

4.3 Mode-dependency due to differential cooperation behaviour

A major strength of interviewer administered surveys possibly lies in their greater persuasive power, because the interviewer is able to convince respondents to comply with a request additional to any written solicitation. Several social-psychological processes, taking place during the interviewer-respondent interaction, are simply absent in self-administered surveys (Groves, Cialdini, & Couper, 1992; Groves & Couper, 1996). If such persuasive power varies over the population, e.g. if older people are more easily persuaded by interviewers, mode-dependent selectivity due to cooperation is caused at least between interviewer- and self-administered modes (Morton-Williams, 1993; Snijkers, Hox, & de Leeuw, 1999; Dijkstra & Smit, 2002).

Another cause of selectivity due to mode-dependent cooperation is the distribution of preferences for certain design characteristics in the population, as implied by leverage-salience theory (Groves, Singer, & Corning, 2000). Leverage-salience theory suggests persons approached with a survey request judge those elements of the request, which are salient to them. If such elements are valued positively, response probability is increased.

The survey mode is a very important design characteristic, obviously. Mode preferences would lead to mode-dependent bias, if they exist and if they are unequally distributed across the population and correlated with other socio-demographic or attitudinal indicators. There is, however, only inconsistent indication that differential preferences do lead to differential mode choice. Diment & Garret-Jones (2007) find that when offering respondents either mail or web, more respondents choose the web mode. Dillman, West, and Clark (1994) found, however, no effects in a similar study comparing telephone and mail modes. Similarly, studies trying to measure respondent mode preferences, have not confirmed behavioral consistency of respondents in later surveys. Rather respondents answered consistently with a preference for the mode, in which they happened to be interviewed about their preference (de Leeuw, 1992).

Expectations about the impact of cooperation. Theoretical expectations are not fully consistent regarding mode-dependent cooperation. On the one hand, survey cooperation can be expected to be very mode-dependent, because there simply might be distinct mode preferences in the population. Under this perspective we would expect various differences across all modes. On the other hand, likewise the argument made for contact, the interviewer plays a very central role in the persuasive process, which misses in the non-interviewer modes. If the persuasiveness by the interviewer varies in the population, there will be a general bias that is introduced for the interviewer modes. Such bias is then absent or different for non-interviewer modes, obviously.

4.4 Sequential mixed-mode designs

If survey response is mode-dependent, the basic assumption of mixed-mode surveys is confirmed. The question remains, however if sequential mixed-mode surveys are successful in making use of differential response. Consequently, our third research question is:

> RQ3: *Do sequential mixed-mode designs produce samples less biased by selection and hence more representative samples than single-mode designs?*

As noted in introducing this paper, ordering modes of administration in sequence often increase overall response rates (Link & Mokdad, 2006; Dillman, Phelps, et al., 2009). The type of mode mix is surely relevant in this respect. Link and Mokdad (2006) found response rate increases when sequencing mail or web with Telephone interviewing, confirmed by studies of Green, Speizer and Wiitala (2005) and Dilman, Phelps et al. (2009). Dilman et al. additionally showed that incentivized mail follow up surveys to Telephone are successful. However, Lynn, Uhrig & Burton (2010) did not find an increase in response rate when combining CAPI and CATI in sequence. We may speculate that the switch from a mode without interviewer to a mode with an interviewer is most successful.

Less attention has been given to sample bias reduction by combination of modes. In a study of health plan participants in the US, Fowler et al. (2002) reported improved sample quality with respect to age and gender distributions in a mail survey with a telephone follow-up. Link and Mokdad (2006) compared web and mail survey with CATI follow-up to regular CATI response and found with respect to race, income and education less bias in the joint samples. However, age distributions were more biased than in the CATI mode alone. In a pre-selected US population sample, Dillman, Phelps et al. (2009) compared multiple mixed-mode strategies and found that no mode sequence was successful in reducing bias on all assessed background characteristics (gender, education, age, income).

Expectations about the impact of sequential mixed-mode designs. We expect that that mixed-mode designs increases expected response rates relative to the first mode used in the sequence when a non-interviewer mode is followed by an interviewer mode. Due to the scope of this paper we limit to these comparisons. Moreover, we expect that modes combined in sequence assist in reduction of absolute sample bias and increase sample representativeness.

4.5 Impact of target variables

Another hypotheses derived from leverage-salience theory is that a survey’s topic, its salience and perception can strongly contribute to overall bias on target variables. As practical situation assume, for example, that a certain state on a target variable makes the survey more sensitive or more interesting to some person, increasing his or her response probability. Topic interest and sensitivity indeed have been shown to be of chief importance for survey response (Groves, Presser, & Dipko, 2004; Voogt, 2004; Groves et al., 2006; Adua & Sharp, 2010).

Any survey therefore is generally suspected to be biased with respect to its target variables. Moreover, topic sensitivity can be suspected to be of higher relevance for cooperation in interviewer administered surveys decreasing cooperation for those respondents with sensitive states regarding the survey topic (de Leeuw & Hox, 2010). This is a situation when we would expect mode-dependent cooperation behaviour traceable for target variables and their correlates. We therefore conclude by asking:

RQ4: *Is there mode-dependent response also on the survey's target variables?*

Noteworthy, RQ4 usually cannot be evaluated in practice, because no validation data are available for any survey's target variables. As illustrated in the next section, our research design allows doing so, however, to some good extent.

5. Method

5.1 Study Design

We administered a 4x2 randomized factorial experiment (N=7,981) with a probability person sample of the general population in The Netherlands, drawn from the population register. Each subject was randomly assigned to any of four survey modes, computer assisted face-face interviews (CAPI), telephone interviews (CATI), paper-based mail interviews (mail), or computer assisted web interviews (web). Information on landline telephone access was available for all respondents based on linked register data. Subjects without a registered telephone number assigned to the CATI condition thus immediately were coverage nonrespondents.

Irrespective of condition, the topic of the survey was announced in a notification letter to be the Dutch Crime Victimisation Survey (CVS), a survey of security and neighbourhood perception, past victimisation and attitudes about the performance of the police. Letters in the mail condition additionally contained a paper questionnaire, whereas letters in the web condition only contained a web link. Subjects in the interviewer administered modes were contacted by trained interviewers to make an appointment for an interview by telephone (in CATI) or face-to-face in subjects' homes (in CAPI). The fieldwork agency and survey sponsor was Statistics Netherlands. An incentive was not administered. In case of nonresponse in the self-administered modes, subjects received up to two mailed reminders. In case of noncontact, multiple call attempts (CATI) were made and CAPI interviewers tried to establish contact at subjects' addresses, if necessary by leaving a postcard with a call-back request. The fieldwork strategy employed is standard for almost all surveys at Statistics Netherlands providing external validity to our design.

The fieldwork covered a time period of 4 weeks. Following this first wave, all subjects, i.e. wave 1 respondents and nonrespondents, were re-approached with a second survey (Wave 2). The second survey mode was administered randomly either in CAPI (75%) or CATI (25%), where respondents without a landline connection were all assigned to the wave 2 CAPI mode. The choice of wave 2 mode was motivated by the need for maximized response and full coverage using CAPI, but it was constrained by cost considerations (resulting in substitution of 25% of units by CATI).

The goal of wave 2 was to collect validation data on target variables of the first wave, recall data of first wave contact, and internet access information of the household. This survey wave was labelled with a different topic, i.e. as a quality assurance survey, and also included some items regarding survey quality not reported on here. Respondents from the first wave were explained in a notification letter that they were approached for a quality assurance of surveys by Statistics Netherland. Nonrespondents were approached as in the first survey wave, but in either CAPI or CATI. Due to the mode-switch to either CAPI or CATI, this design allowed an assessment of sequential mixed-mode response for wave 1 web or mail nonrespondents.

5.2 Operationalisation

We will assess mode-dependency based on two sets of variables, standard socio-demographics and target variables of the CVS (Table 1). One limitation of our data base was information on persons' educational level, which is not part of the Dutch population register.

To measure, whether contact was established in the self-administered modes, we added a recall question inquiring wave 2 respondents, whether subjects could remember that they had received the invitation letter sent prior to wave 1. We use this recall question as an analogue to the interviewer observed contact data available for the interviewer modes. Recall surely does not equal actual contact, but might assist in giving useful estimates within the boundaries of some memory errors.

Table 1: Overview on variables used for bias assessment and origin (register / wave 2)

Socio-demographics (Register)	*Target variable (Wave 2)*
Gender Age Income Ethnicity Marital Status Household size Children in the household Degree of Urbanization Inhabitant of a big city	*Subset of CVS key variables:* Social Neighbourhood rating Neighbourhood problem rating Past Victimization (summaries) Police contact within past year Household Internet Access Recall of the letter sent prior to wave 1

Web coverage generally is subject to definitional debate (Couper et al., 2007; Blyth, 2008), because household internet coverage is not a necessity for web survey participation. After all, a web survey can be completed elsewhere than at home by the non-covered respondent. We decided to restrict to in-house internet access. We inquired whether respondents had a desktop or notebook computer at home, with which they can access the internet. We restrict to this definition excluding handheld and other mobile devices, because the particular web survey could only be accessed using a classical browser system and would not display properly on handhelds.

There are to our knowledge no prior findings on whether non-availability of in-house internet access prevents participation in web surveys. As it turned out, only four persons within the wave 1 web condition responded despite indicating in wave 2 that there was no internet access in their household. This supports the validity of our definition and sequencing in Figure 1.

5.3 Implementation

5.3.1 Fieldwork and response

Fieldwork took place in two phases (wave 1 and 2) between March and June 2011. Table 2 illustrates how response to the first and second wave was structured for all modes. Due to interviewer reports, we know about non-contact and refusal in CAPI and CATI. Non-contact was much higher in CATI (9.2%) than CAPI (5.7%), though overall low. Much more significant was non-cooperation, being lower in CATI, however, than CAPI due to a lower percentage of refusal and higher propensity to successfully make appointments in CATI. Overall response rates fulfilled general expectations regarding mode-dependent marginal survey response: CAPI (64.8%) and CATI (67.5%) were strongest followed by mail (49.8%) and web (28.7%; AAPOR RR1). Note, however, that 28.5% of the CATI gross sample is lost due to non-coverage. The percentage completed of all sampled units drops to 45.1%, if this is taken into account, which is slightly lower than mail response.

During wave 1 another share of respondents was identified as coverage non-respondents, e.g. in cases when contrary to our information a landline account was disconnected. These nonrespondents could not be approached again in wave 2 due to internal administrative reasons. For the interviewer modes we also know if respondents were not able to participate, a generally low percentage (e.g. due to language deficits or sickness). This information is not available for the non-interviewer modes.

For the second survey wave respondents were randomly assigned either to CAPI (75%) or CATI (25%), where all non-landline households were assigned to CAPI. Overall, cooperation was slightly worse in the second wave in both modes yielding overall lower response.

5.3.2 Validity of wave 2 measurements: contact recall and web coverage

The validity of contact assessment in the self-administered modes strongly depends on wave 1 (non)respondents ability to recall. Therefore, recall of wave 1 contact certainly was subject to some measurement error, whose size, however, can be assessed and corrected at least partly for ‘false negatives’. That is, wave 1 respondents, who falsely replied in wave two that they could not recall the solicitation letter, which they must have received and read to participate. We correct the recall measure for these mistakes.

Table 2: Response to the first and second wave (% of total and AAPOR response rates)

	Wave 1				Wave 2	
	CAPI	CATI	Mail	Web	CAPI	CATI
	%	%	%	%	%	%
Non-eligible	4.6	6.3	0.8	0	6.5	12.2
Wave 1 Not re-approachable[c]					4.7[c]	
Wave 2 Administration error[d]					1.2[d]	
Non-coverage (register)	0	22.1	0	n/a	0	0
Newly identified (fieldwork)	-	4.6[d]	-	-	-	5.1
Non-contact	5.4	6.2	n/a	n/a	5.8	11.2
Refusal	19.8	10.5	n/a	n/a	26.7	22.7
No appointment	5.5	1.2	n/a	n/a	5.9	2.7
Not able[a]	2.9	4.0	n/a	n/a	2.2	3.7
Nonresponse (mail/web only)	-	-	49.8	71.3	-	-
Complete (partial)	61.8 (.5)	45.1	49.4	28.7	54.6 (0.4)	51.0
Total (N=8,781)	2,182	2,200	2,200	2,199	6,404[e]	1,620[e]
Coverage Rate[b]	1	71.5	1	n/a	1	94.6
AAPOR Contact Rate	94.3	90.8	n/a	n/a	93.9	87.7
AAPOR Cooperation Rate 2	68.7	74.3	n/a	n/a	61.1	63.7
AAPOR Response Rate 2	64.8	67.5	49.8	28.7	57.4	55.8
% Complete of Total	61.8	45.1	49.4	28.7	54.6[e]	51.0[e]
% Complete of Eligible	64.3	48.2	49.8	28.7		

a.: e.g. long term illness, deceased, language problems etc.; following AAPOR definitions, 'not-able' can only be determined conditional on contact in CAPI/CATI

b.: Computed as Covered/(Total-Noneligible))

c.: A group of 415 persons (4.7%) was not approached again, because interviewers noted in wave 1 that they would feel uncomfortable with a repeated solicitation.

d.: During wave 1 it turned out that a group of 101 units classified as accessible by landline telephone, in fact was not. Due to internal administration this group was classified as ineligible to wave 1 and therefore was also ineligible to wave 2 (classified as wave 2 administration error).

e.: Total minus 'Wave 1 not re-approachable' minus 'Wave 2 Administration Error'

Following adjusted recall, 11.6% of wave 2 respondents in wave 1 'mail' (22.4% before adjustment) and 21.5% (28.2%) in wave 1 'web' are classified as non-contacts (data including persons without web access). Considering table 2 we can see that noncontact rates for wave 1 CAPI (5.7%) and CATI (9.2%) respondents are in the same magnitude, but that non-recall rates are higher for mail and web. Since we cannot adjust for 'false positives' (wave 1 nonrespondents with false recall), recall rates are rather over- than underestimations. The adjusted recall measure will be used in the following.

Web coverage was measured for all wave 2 respondents (n=4297), regardless of wave 1 mode. The estimated web coverage rate is 90.0%, which is very close to the population estimate available from national statistics (94%; Statistics Netherlands, 2011).

5.4 Statistical modelling

Let J be a categorical variable with K classes, then we have for individual I, the '*mode-dependence model*', subject to identification constraints:

$$P_m(C_{M=m,i}=1 \mid X_j, M=m) = link(\beta_{C,m,j,k=1} X_{k=1,i} + \ldots + \beta_{C,m,j,k=K} X_{k=K,i})^{-1} \qquad (7)$$

Model (7) can be used to assess mode-dependence definition 2. In principle six separate mode comparisons would be possible. Since there are 24 different classes over all covariates, we could test 144 parameter differences. Chance capitalization turns results from such a large amount of (in)dependent tests hard to interpret. We, therefore, use a more restrictive and powerful test based on a simultaneous hypothesis:

$$H_0 : \beta_{C,M=1,j,k=1} = \ldots = \beta_{C,M=1,j,k=K} = \beta_{C,M=2,j,k=K} = \ldots = \beta_{C,M=M,j,k=K} \qquad (8)$$

That is, we impose a simultaneous equality constraint on all mode-dependency indicators for a given covariate *J* and a given response type *C*. The hypothesized *'independence model'* thus is

$$P_m(C_{M=m,i}=1 \mid X_j, M=m) = link(\beta_{C,m,j,k=1} X_{k=1,i} + \ldots + \beta_{C,m,j,k=K} X_{k=K,i})^{-1} \qquad (9)$$

One way of testing H0 is to specify the dependence model as interaction model using a mode indicator. This model contains all interactions of categories of X and mode, subject to the inclusion of reference categories for identification. We, then, use a log-likelihood ratio test (LRT) of the independence model against the dependence model (for LRT see Agresti, 2002, p. 141-142). Our final model will test H0 for a given category *j* of *X* while keeping constant all other interactions in the full and the independence model. If H0 is rejected for J we speak of mode-dependency of response and nonresponse bias on *X* following definition 2.

To assess mode-representativeness (definition 3), we estimate R-indicators for each of the four conditional versions of the dependence model. We use analytic standard error and confidence interval estimates discussed in Shlomo, Skinner, and Schouten (2012).

5.5 Summary of design-inherent assumptions

Since socio-demographic indicators are taken from population registers, the analyses based on these variables will not depend on the second survey wave and can be generalized to the population. However, if the models estimated based on the second wave data shall be generalized to the population, we would need the target variables, web access and recall data to be unrelated to the response process. This assumption seems plausible for web access and recall data.

However, the target variables may not only determine response to wave 1, but also response to wave 2 weakening our ability to extrapolate to the population. Nevertheless, we can consider this analysis valid relative to the implementation of a regular CAPI/CATI survey. Thus we can say that mode-dependent bias in web, mail, and CATI of wave 1 is assessed relative to the regular response in a CAPI/CATI survey. Wave 2 response, therefore, should not be affected by the mode of wave 1 in any way and it should closely resemble a regular CAPI /CATI response.

This implies, first, that there was no systematic attrition between waves, which is typical in panel-like surveys. We provide analysis that these assumptions are plausible in Appendix A. Second, measurements obtained in wave 2 are assumed to be independent of participation in wave 1 and constant across the time of the study period. Since the second survey was administered 4 to 8 weeks after the first survey, we argue that the repeated measurements are not affected by (non)participation, but that strong change is unlikely.

Noteworthy, by using single-mode measurements from wave 2 we avoid the usual problem of confounding effects of measurement and selection in experimental designs, which was explained in section 3.

6. Results

In this section we will address each research question and related expectations, respectively.

6.1 Mode-dependency of the outcome of the response process

Assessment based on definition 1 (mode-dependent expected response rates). In section 5.3 we already described the response rates obtained for each of the mode conditions. Generally, we find that consistent with expectations the web condition yielded by far the lowest response rate (29%). CAPI yielded the highest response rate (65%), while CATI (45.1%) and mail (49.1%) conditions were in between. Differences are overall significant ($p<.0001$). CAPI response was significantly higher, web response significantly lower than in all other mode conditions. CATI and mail response rates did not differ significantly. These results meet expectations regarding the rank of web and CAPI conditions in the order of expected response rates. Mail response was surprisingly high, however.

Assessment based on definition 2 (mode-dependent relative sample bias). Table 3 shows test results for main effects of socio-demographic indicators on overall outcome of the response process. Loglikelihood-ratio-tests are used to test the hypothesis that all dummy indicators of a given categorical covariate are zero. Due to the large number of dependent tests (32 in total), we apply a Bonferroni correction to overall .05 type-1 error probability. This yields a significance level of .0015 for all tests (indicated by bold starred p-values). As table 3 merely shows p-values of statistical tests, we will mention important contrasts between groups in the text. The interested reader may want to consider Appendix B for detailed model estimates.

Consider the ‘absolute bias’ section of table 3. Nonresponse bias is detected in all modes in a quite distinct pattern. While, for example, the CAPI and web mode are unbiased with regard to the ‘gender’ distributions, CATI and mail are biased (women overrepresented). The web is fully unbiased with regard to the ‘age’ distribution, whereas the other modes are biased in some way. Similar differences are found for ‘income’, ‘Marital Status’, and ‘household size’. Only the ethnicity indicator shows bias for all groups (especially non-Western minorities are underrepresented). Regional bias is found for the CAPI mode as indicated by ‘Degree of Urbanization’ and ‘agglomeration’, somewhat less certain due to higher p-values. Also note that the web condition only shows a strong bias for the income distribution (higher incomes are strongly overrepresented in the sample), whereas the other bias estimates for web are less certain due to higher p-values.

Table 3: Main effects of socio-demographic predictors on absolute nonresponse bias (p-values<.01 [single-mode], <.05 [interaction] from LRT, bold starred indicates significance after Bonferroni correction)

	Absolute bias				**Interaction Tests (Relative Bias)**			
	CAPI	**CATI**	**Mail**	**Web**	**Overall**	**Mail vs. Web**	**CAPI vs. CATI**	**Interviewer vs. Nonintervie wer**
Gender	-	**.0005***	**<.0000***	-	**<.0000***	**.0005***	**.0014***	-
Age	**.0002***	**.0016***	.0045	-	-	-	-	.0086
Income	-	-	**<.0000***	**<.0000***	.0361	-	-	.0063
Ethnicity	**<.0000***	**<.0000***	**<.0000***	.0038	.0077	-	-	.0049
Marital Status	-	.0056	**.0009***	-	-	-	-	-
HH size	-	**.0003***	-	-	-	-	-	.0389
Urbanization	-	-	-	-	-	-	-	.0493
Agglomeration	.0036	-	-	-	-	-	-	-
N	2081	2062	2182	2199	8524	4381	4143	8524
McFadden R^2	.055	.079	.068	.052	.109	.094	.085	.074
R-Indicator	.765 (.021)	.697 (.019)	.717 (.019)	.797 (.020)	-	-	-	-

Only the ethnicity indicator shows bias for all groups (especially non-Western minorities are underrepresented). Regional bias is found for the CAPI mode as indicated by 'Degree of Urbanization' and 'agglomeration', somewhat less certain due to higher p-values. Also note that the web condition only shows a strong bias for the income distribution (higher incomes are strongly overrepresented in the sample), whereas the other bias estimates for web are less certain due to higher p-values.

Now, we examine how these effects differ across mode conditions (Definition 2). Consider the interaction tests column in table 3. Conditional response probabilities differ especially on the gender stratum, and with less certainty regarding income and ethnicity. Considering the many biases reported, we were surprised that only few contrasts reach significant levels.

To analyse differences more thoroughly, we consider comparisons of CAPI/CATI and mail/web respectively. For this analysis we look only at CAPI/CATI and mail/web interactions in a common model. Here we do not find any significant differences, except for gender variable. The origin of the gender difference lies in the bias of mail (odds of women are 1.674 of the odds for male persons to respond) and CATI conditions (odds ratio of 1.429), while web and CAPI are unbiased.

In face of such remarkable homogeneity, we decided to pool the interviewer and non-interviewer modes respectively. We test, whether there are differences in nonresponse outcome between the pooled conditions. Obviously, pooling now dilutes the differences on the gender variable, whilst revealing a series of differences that persist between the pooled interviewer and self-administered modes. Though all of these differences do not reach the

Bonferroni adjusted significance level, we believe that such clustered and cumulative evidence, points to the fact that mode-dependent response is present. Differences are probably present on age, income and Ethnicity distributions ($p<.01$) and possibly also on household size and urbanization indicators ($p<.05$). The initial analysis in column 'Overall' seems to attenuate these differences, possibly due to too low power, which is increased by the pooling of the modes. Thus, contrasts appear more pronounced in the analysis.

It is instructive to consider the signs of effect estimates not shown in table 3, to understand the homogeneity and divide between pooled conditions. Regarding *age*, a strong diametrical bias is found. Whereas in the CAPI/CATI modes, the age groups between 25 and 54 are underrepresented, bias is absent in these groups in web/mail. However, the mail condition (not web), contains over-proportionally many respondents aged 55 or above. Concerning *income*, we find more biased samples in the mail and web modes, only the CAPI/CATI modes are rather unbiased. Here the income brackets over 30,000 Euro are over-represented in web and mail. Pooled modes also differ strongly on the *ethnicity* of samples. Whereas non-western foreigners are equally under-represented in all modes, only the interviewer modes are biased downwards for western foreigners.

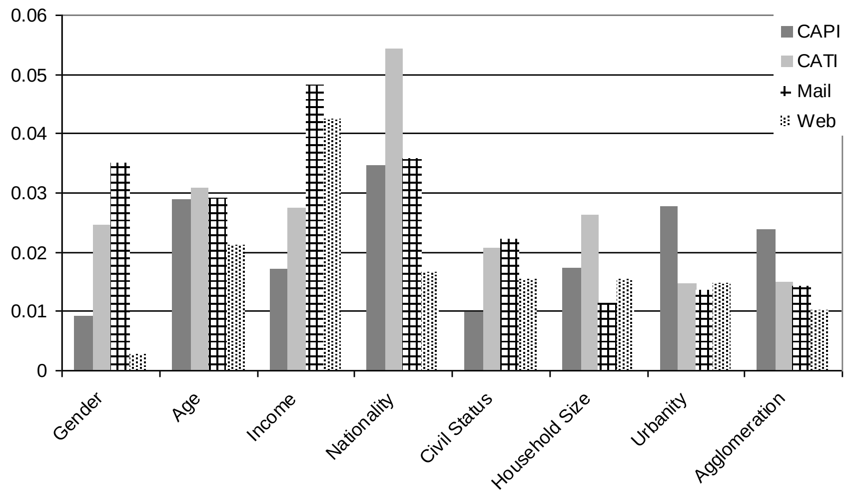
As it becomes clear from these details, absolute bias is similar for CAPI/CATI and mail/web, respectively, but distinct when comparing pooled modes.

Assessment based on definition 3 (mode-representativeness). R-indicator estimates are provided in table 3 for each mode condition. First of all, we note that the four indicators are not very widely spread suggesting no huge differences in representativeness. CATI yields the lowest R-indicator (.697), whilst web yields the highest (.797), i.e. none of the modes is fully representative. Taking into account the variability of the R-indicators estimating 95% confidence intervals, we may note that the CATI and mail (.717) conditions probably yield worse representativeness estimates than the web mode. CAPI (.765) and CATI are also significantly different, while mail and CATI are not. A CAPI-mail comparison also does not exceed sampling variation.

Certainly, the low spread of indicators matches the finding based on the overall tests (Table 3): overall mode-dependency is not strong. However, the web and the mail mode response probability variances are significantly different. We inspected conditional partial R-indicators (Schouten et al, 2011) to assess the reasons for the difference (Figure 2). Partial R-indicators are measures of covariates' individual contribution to overall representativeness. The higher a partial R-indicator, the higher is the loss in representativeness. Whilst web and mail are very homogenous on all partial R-indicators, we find a strong difference due to gender (and less pronounced due to ethnicity). Regarding CAPI and CATI we find a similar situation, i.e. differences due to gender and ethnicity.

The difference noted above regarding gender (cf. table 3) contributes to a major loss in representativeness of mail against web, which compared amongst each other are rather similar on all other covariates.

Figure 2: Conditional partial R-indicators for socio-demographic covariates



6.2 The sources of mode-dependent survey response

Our second research question asks how mode differences are created in the stepwise process of survey response.

Assessment based on definition 1. Table 4 shows response rates conditional on success on the previous response stage. Coverage and contact rates for mail and web modes are estimated based on wave 2 response. Since wave 2 respondents generally had a higher response probability in wave 1 all rates are overestimations. We adjusted for this using post-stratification, where strata are represented by wave 1 (non-) response groups.

It is obvious, that CAPI and mail have full population coverage, whereas CATI and web suffer from under-coverage problems. The estimate of internet access based on subjects in the web condition is 89.1%: significantly higher than CATI response. 28.5% of the CATI gross sample is lost due to non-coverage.

Contact was based on interviewer reports for CAPI and CATI. For the interviewer modes, we find slightly higher non-contact and non-cooperation in CAPI than in CATI. Non-contact was much higher in CATI (9.2%) than CAPI (5.7%), though overall low. Obviously non-cooperation had by far stronger impact on the final response rate than contact.

Surprisingly, Web contact recall was much lower than mail recall. Moreover, cooperation rates (calibrated to overall response rates) must have been far lower in web than mail (estimated 45 vs. 60 percent). Comparing contact rates across all modes we find that the mail mode is close to CAPI and CATI, whereas contact in Web is lowest. Regarding cooperation Mail is fairly close to the rate obtained in CAPI, whereas Web certainly yields lowest cooperation of all modes.

Table 4: Coverage, contact (conditional on coverage), cooperation (conditional on contact) and response rates (% complete of eligible) by mode conditions (in %)

	CAPI	*CATI*	*Mail*	*Web*

	(%)	(%)	(%)		(%)	
			Unweighted	*Weighted[a]*	*Unweighted*	*Weighted[a]*
Coverage	100	71.5 (n=2062)	100	100	89.1 (n=1091)	87.1
Contact (recall)	94.3 (n=2081)	90.8 (n=1474)	88.4 (n=1112)	82.9	79.9 (n=972)	73.9
Cooperation	68.2 (n=1963)	74.2 (n=1338)	74.6 (n=983)	60.0	56.6 (n=777)	44.6
Response rate	64.3 (n=2081)	48.2 (n=2062)		49.8 (n=2182)		28.7 (n=2199)

a. Wave 2 contact recall estimate weighted to wave 1 response strata. Cooperation rates calibrated to overall response rate.

Assessment based on definition 2. Given the divide that we found mainly between the pooled mode conditions, our first approach is studying the source of these differences. The first source is represented by differential coverage between web and CATI. Table 5 shows *p*-values of covariates modelling household coverage (columns 1 and 2). Both telephone and web populations are quite distinct from non-covered populations but in different ways: probabilities for web coverage strongly decrease in higher age groups (e.g. the group of 65+ e.g. has an odds ratio of .008 vs. 14-25 year old respondents to be covered by web), for lower income and in 1-person households. Probability for telephone coverage, to the contrary, is highest for older persons and increases in non-urban regions. By this process, telephone and web introduce differential coverage bias, especially on the age and income distributions (Table 6, column 3). Web populations are relatively stronger biased for younger and more affluent persons, telephone populations are older but unbiased for income. In the final outcome, mode-dependent bias is found in exactly these directions.

However, there are differences that cannot be explained by coverage alone. The second reason for the pronounced differences, therefore, is differential contact and cooperation behaviour. The differences are most apparent when considering the cumulative effects of contact and cooperation conditional on coverage (table 6, column 8), as compared to conditional contact and cooperation (columns 5 and 6). Here, a much wider ranger of differences emerges than for coverage. Pooled self-administered conditions are more biased for women, older age groups, higher incomes, and higher response propensity in large cities. They are unbiased with respect to western foreigners, whereas in the pooled interviewer conditions this group is underrepresented.

We may summarize that bias between pooled mode groups is apparent on all stages of response.

Table 5: Sources of mode-dependent nonresponse between pooled interview and non-interviewer modes (p-values<.05 from LRT, bold starred indicates significance after Bonferroni correction)

	Coverage (Absolute Bias)		Coverage (Relative bias/Interaction Tests)		Pooled modes (Relative bias)		Outcome (Relative Bias for pooled modes)	
	CATI (1)	Web (2)	CATI vs. Web (3)	Pooled mode groups (4)	Conditional Contact (recall) (5)	Conditional Cooperation (6)	Unconditional (7)	Conditional on coverage (8)
Gender	**.0028***	-	-	-	-	-	-	.0063
Age	**<.0000***	**<.0000***	**<.0000***	**<.0000***	**.0001**	.0021	.0086	.0067
Income	-	**<.0000***	**<.0000***	**<.0000***	-	-	.0063	.0088
Ethnicity	**<.0000***	**<.0000***	-	-	-	-	.0049	.0473
Marital Status	-	-	-	-	-	-	-	-
HH size	**<.0001***	**<.0000***	**.0146**	**-**	**.0008**	-	.0389	.047
Urbanization	**<.0037***	-	-	-	-	-	.0493	-
Agglomeration	-	-	-	-	-	.0228	-	.0299
N	2062	4296	3153	7416	5639	5061	8524	7817
Pseudo R^2	109	.366	.215	.161	.090	.039	.074	.091

Having made these statements about the impact of coverage, the reader may wonder why we did find overall homogeneity for the interviewer and self-administered modes, respectively (cf. table 3). The coverage bias could have caused also differences between CATI and CAPI and web and mail, of course. We found that the differences due to non-coverage of telephone and web are mitigated in the process of cooperation and contact. Table 6 compares the contact and cooperation stages of the CAPI mode with impact of non-coverage in CATI. One can notice that the structure is remarkably similar: we find age, ethnicity, household size, and Degree of Urbanization biased by contact in CAPI, which just happens to be the coverage deficits in CATI. A more qualitative inspection of the particular parameter estimates gives further ground to this argument. CAPI is very good at contacting old persons (65+), which just is a group that above average holds landline connections. CAPI also looses many western foreigners during the contact stage, which is also the case for telephone coverage. Finally, larger households are more easily contacted in CAPI, but also hold more landline connections. A similar argument can be made for the web/mail comparison.

The mail mode seems to produce at least a part of the web coverage bias in the process of contact and cooperation. For example, do more affluent respondents cooperate more frequently in mail, which are just those respondents more frequently covered by web. Moreover, comparing the coverage effects of web on age and household size with the bias outcome for web (table 3) we find that these biases are mitigated to resemble mail response more closely.

Finally, table 6 illustrates the relative biases between CAPI and CATI as well as mail and web. One may confirm that the overall homogeneity is also found for contact and cooperation stages (columns 12 and 13; 17 and 18).

Assessment based on definition 3.

Figure 3a (first panel) shows the R-indicators estimated for the four modes in a direct comparison of cooperation, contact and coverage. It is apparent that non-cooperation and non-contact reduces representativeness in all modes, but in a very similar manner. The loss in representativeness is larger due to cooperation than due to contact. Moreover, the mail condition looses least due to contact (recall).

Table 6: Mode-dependent bias (absolute/relative) on coverage, contact (conditional on coverage) and cooperation stages (conditional on contact; p-values<.05 from LRT, bold starred indicates significance after Bonferroni correction)

	Absolute Biases CAPI or CATI			Relative Biases CAPI vs. CATI		Absolute Biases Mail or Web			Relative Biases Mail vs. Web	
	Coverage (CATI)	Contact (CAPI)	Cooperation (CAPI)	Contact	Cooperation	Coverage (Web)	Contact (Mail)	Cooperation (Mail)	Contact	Cooperation
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Gender	**.0028***	-	-	-	.0020	-	.0478	.0060	.0301	.0415
Age	**<.0000***	**<.0000***	.0068	-	-	**<.0000***	-	.0309	-	-
Income	-	-	-	-	-	**<.0000***	-	.0268	-	-
Ethnicity	**<.0000***	.0109	**<.0000***	-	-	**<.0000***	.0489	.0019	-	-
Marital Status	-	-	-	-	-	-	.0352	-	-	-
HH size	**<.0001***	**<.0000***	-	-	-	**<.0000***	**.0026***	-	-	-
Urbanization	**<.0037***	-	-	-	.0062	-	-	-	.0075	-
Agglomeration	-	.0035	-	-	-	-	-	-	-	-
N	2062	2081	1963	3555	3301	4296	1112	983	2084	1716
Pseudo R^2	.109	.183	.039	.140	.044	.366	.067	.078	.068	.086

The impact of non-coverage on representativeness is much stronger. Internet non-coverage reduces the sample representativeness to as low as R=.667. However, the final response sample of the web condition yielded an R of .797, which is significantly higher than the sample representativeness after coverage. One understands this by noticing that representativeness evolves in the course of survey response (Figure 3b). Contrary to figure 3a, R is computed here cumulatively, after coverage, after contact (including non-coverage), and after cooperation (including noncontact). Clearly, CAPI and mail modes, start with *R=1*, because non-coverage is absent in these modes. As noted, the loss in representativeness due to noncoverage in both CATI and web is high. On the contact stage the representativeness of all modes decreases, but we may note that the mail mode does significantly better than CAPI. As was noted above, it hardly looses in the process of contact. While CATI also looses about equally strong as CAPI, now being at .654 (C2), an interesting development regarding the web mode starts. The R-indicator for web increases to .733 (Web, C2). Finally, cooperation brings down CAPI representativeness to .765 (C3), while CATI response stays comparatively low (.697). The loss of representativeness due to non-cooperation in the mail mode, however, is extremely strong (.717). The surprising trend regarding web response continues. After taking into account selectivity due to cooperation web reaches R=.797 (Web C3). Web thus yields the highest R of all modes.

Figure 3a: Mode-representativeness with 95% confidence intervals per response stage

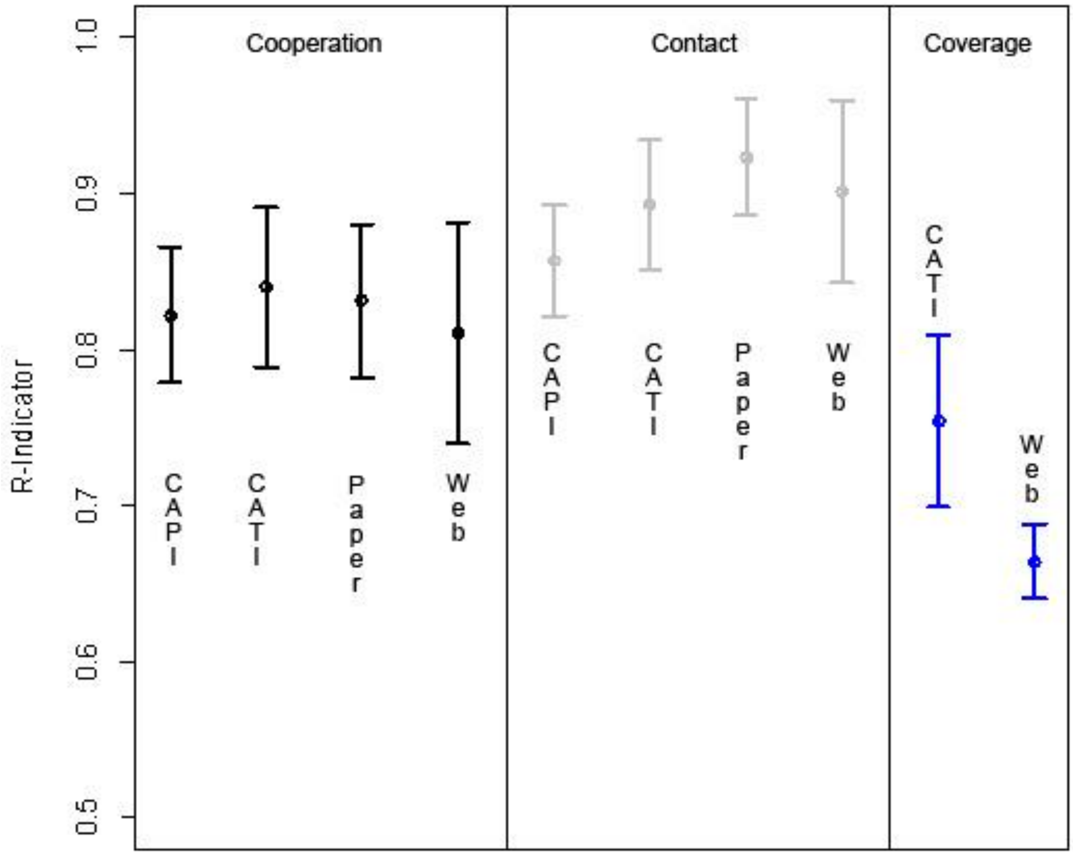
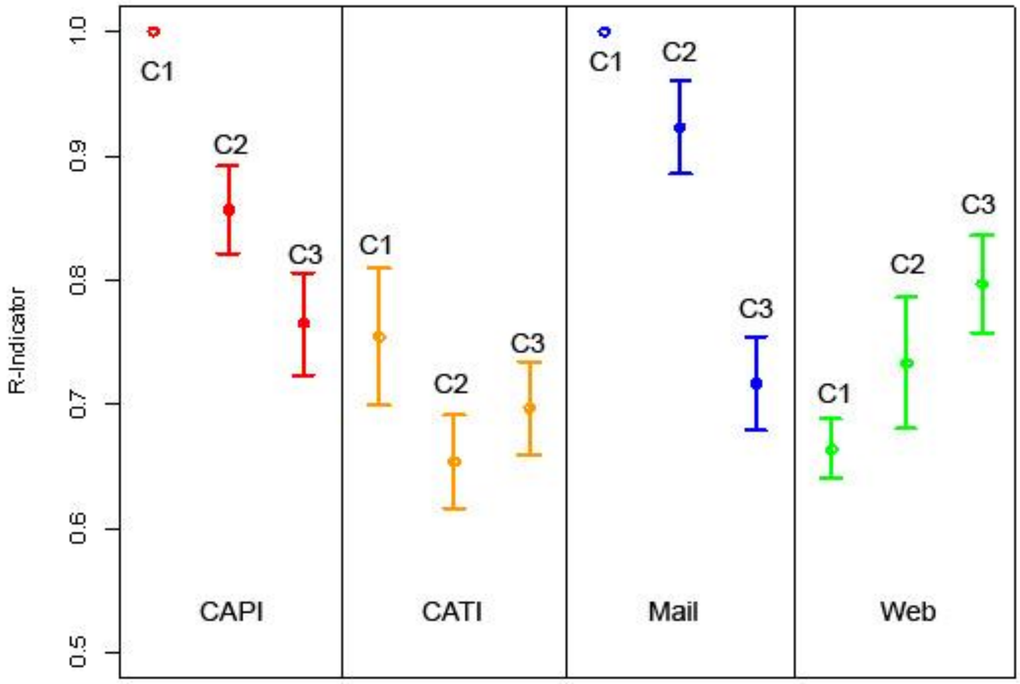


Figure 3b: Cumulative R-indicators with 95% confidence intervals per mode and response stage (C1: coverage, C2: contact; C3: cooperation)



From this analysis it becomes clear that deficits due to coverage need not mean a systematic loss in representativeness. In fact, CAPI and mail modes incur losses in representativeness due to contactability (CAPI) and cooperation (mail). These are less dramatic for CATI and web. In fact, for the web mode the process of contact and cooperation balances the response distribution to an extent that makes web overall the most representative mode.

6.3 The impact of sequential mixed-mode designs

Our third research question concerned the impact of sequential mixed-mode designs on sample bias. We used the second survey wave as a sequential expansion to the self-administered modes of wave 1 (mail or web). We consider either second wave CAPI alone or CAPI and CATI jointly (CAPI+CATI). The response sample size to CATI did not allow a separate analysis. In terms of response rate the sequential designs boost web response from 28.7% to 59.9% adding CAPI (59.1% CAPI+CATI). Sequencing mail with CAPI increased response from 49.8% to 67.6% (67.9% CAPI+CATI). This matches general expectations we had about sequential mixed-mode designs.

Consider table 7 illustrating the main effects of variables predicting response for the sequential designs. Web response was biased for Marital Status, household size and ethnicity, while the sequential mixed-mode is balanced for these characteristics. However, the income bias in web (more affluent persons are overrepresented), which is mainly due to coverage properties of the internet, is not reduced. Moreover the sequential design is biased for urban agglomeration indicators. Considering the R-indicators for this sequential design, we note that although the web appears more balanced based on significance test, these changes only marginally improve overall representativeness. The latter is undermined by strong overall bias of the income distribution in the web condition.

Table 7: Main effects of socio-demographic predictors on overall response in sequential designs (p-values<.05 from LRT, bold starred indicates significance after Bonferroni correction)

	Sequential Mixed-Mode (Absolute Bias)				*Single mode (Absolute Bias)*	
	Web-CAPI	*Web-CAPI+CATI*	*Mail-CAPI*	*Mail-CAPI+CATI*	*Mail*	*Web*
Gender	-	-	**.0005**	**.0002**	**<.0000***	-
Age	-	-	-	-	.0045	-
Income	.0160	.0066	**<.0000**	**<.0000**	**<.0000***	**<.0000***
Ethnicity	-	-	.0202	.0068	**<.0000***	.0038
Marital Status	-	-	.0309	.0375	**.0009***	.0139
HH size	-	.0195	-	-	-	.0110
Degree of	-	-	-	-	-	-
Agglomeration	**.0001**	**.0003**	.0400	.0404	-	-
N	1689	2132	1716	2128	2182	2199
Pseudo R^2	.040	.027	.045	.044	.068	.052
R-indicator	.806	.805	.725	.720	.717	.797
	(.023)	(.020)	(.022)	(.020)	(.019)	(.020)

Regarding the mail+CAPI sequential design, a similar conclusion can be drawn. The sample bias of the mail mode cannot be overcome by the net-increase in response and R-indicators only marginally change with respect to the singe-mode mail design. Finally, adding CATI response to the mix does not alter these conclusions. Rather bias is deteriorated by such proceeding.

One can illustrate why sequential mixed-mode cannot compensate for all sample biases using some examples. Consider, for example, the bias on the income variable in mail and web. Here, the CAPI produces a rather unbiased sample. CAPI would need to under-represent higher incomes to compensate for the deficits of web and mail, which simply is not the case. A similar point can be made for the gender bias in mail. Women are overrepresented in mail, whilst unbiased in CAPI, thus compensation is unsuccessful. However, CAPI is biased for urban agglomeration indicators. The sequential mixed mode, therefore, also suffers this bias, though mail and web were actually unbiased in this respect.

6.4 Mode-dependency assessment based on target variables

Our fourth research question asked whether there was mode-dependent bias also on the target variables of the Crime Victimisation Survey. We use the second wave (single-mode) re-measurements of the target variables in the assessment of absolute and relative sample bias (Table 8). Units not responding to wave 2 are deleted from this analysis. Many of the target variables re-measured in wave 2 indicate a different response process in wave 1 (main effects). Especially, the aggregated victimization variables all reach highly significant levels.

Table 8: Mode dependent survey response assessed by wave 2 measurement of target variable, wave 2 modes (CAPI/CATI) pooled (p -values<.05 from LRT, bold starred indicates significance after Bonferroni correction)

	Main effect	Overall Modes	Interviewer-/ Self-administeredpooled
Items social quality neighbourhood:			
State of roads, walkways, and squares	n.s.	-	-
Good playgrounds for children	.0038	-	-
Good provisions for younger persons	n.s.	-	-
People know each other well	.0189	-	-
People treat each other well	.0021	-	-
Nice neighbourhood with cohesion	n.s.	-	-
Feel at home with people	**.0017**	-	-
Have a lot of contact with people	.0143	-	-
Satisfied with population composition	**<.0009**	-	-
Reported scale score	**.0013**	-	-
Items problems in neighbourhood:			
Plastering walls and/or buildings	n.s.	-	-
Harassment by groups of young persons	.0414	-	-
Drunken people on the streets	n.s.	-	-
Unpleasant people on the streets	**.0036**	-	-
Junk on the streets	.0084	-	-
Dog excrements on the streets	**<.0001**	-	-
Destruction of telephone cells, tram or bus stops	.0279	-	-
Drug problems	n.s.	-	-
Reported scale score	.0203	-	-
Victimisation yes/no (past 12 mth.)	**.0013**	-	-
Victimisation violence yes/no (past 12 mth.)	**.0008**	-	-
Victimisation total number (past 5 years)	**.0004**	-	-
Victimisation violence total number (past 12 mth.)	**<.0001**	-	-
Safety Feeling: insecure	n.s.	.0043	.0011
Quality of Life rating (neighbourhood)	**<.0001**	-	-
Police contact (past 12 mth.)	n.s.	-	-
	N=4324	N=4324	N=4324

It is important to note that this ‘nonresponse bias’ can only be extrapolated to the population under very strong assumptions (e.g. response independence from the target variables in wave 2). This analysis, however, gives some idea on the potential for ‘nonresponse bias’ on many target variables, at least relative to the wave 2 survey.

While the interpretability of the main effects is limited, we should be able to detect mode dependent response dependency. If conditional response probabilities for Cati, mail, and web differed in wave 1, this is likely to show in the estimated interaction between the wave 1 mode indicator and the re-measured target variables. It is quite clear from table 5, that for nearly all target variables assessed, we find no relationship at all.

There is one exception, however, regarding the variable ‘safety feeling’. This question inquired whether persons sometimes feel unsafe in their neighbourhood. This relationship could be clearly found by chance. But it is intriguing to note that there is a bias between the interviewer and self-administered modes. In fact, respondents sharing stronger unsafety feelings cooperate less frequently in the interviewer administered modes. Moreover, detailed analyses on the response process showed that this bias is not created by non-cooperation, but by a lower contact rate by interviewers in the group with stronger unsafety feelings.

One alternative explanation for finding mainly insignificant differences, however, is the sample size. In principle it is possible that by increasing sample sizes one could potentially identify smaller effects of selectivity. However, we are confident that we would have identified stronger effects with the give sample size, which in the boundaries of social science research cannot be considered small. We come back to this point in the discussion.

7. Conclusion

Nonresponse problems in face of high costs are urging researchers to adopt alternative designs, such as mixed-mode surveys. Mixed-mode surveys make use of mode-dependent response processes to produce less biased samples. The existence of mode-dependent response processes is, however, rather a common belief than a tested hypothesis. Based on a large-scale national experiment in The Netherlands, we evaluated, whether this belief is reasonable, where mode-differences in response emerge, and if sequential designs help in the reduction of absolute nonresponse bias. In doing so we exclusively focussed effects of selection by differences in the response process. Another important source of error is represented by measurement errors. These are addressed elsewhere in research by Statistics Netherlands (Schouten et al. 2012).

We did so using register type socio-demographics as auxiliary information, because this is the only information available on gross sample level. Furthermore, based on a second survey wave, we could measure sample coverage with internet access as well as recall of wave 1 contact for mail and web. The second survey wave was free of coverage bias being either administered CAPI or CATI, where non-telephone households were interviewed in CAPI. Our coverage estimate was close to the estimate of the regular ICT CATI survey carried out by Statistics Netherlands. The second wave simultaneously acted as a sequential extension of the wave 1 web and mail response samples.

Three different ways to describe mode-dependent response processes are available (Table 9). All are based on nonresponse bias of a population mean (or total) estimate. Those are, mode differences in the response rate, relative bias, and representativeness. Conclusions based on any of the definitions need not match. In particular the response rate is a weak indicator of sample bias. This was also confirmed in our study.

Table 9: Summary of major findings on the mode-dependency of survey response and nonresponse bias

Research Question	*Response rates*	*Sample Bias tests*	*Representativeness*
Overall	CAPI highest Web lowest Mail response surprisingly high	Homogeneity CAPI/CATI and web/mail (except gender) Many differences between pooled modes	Web and CAPI highest, mail and CATI lowest
Major Sources	CAPI: cooperation CATI: coverage, cooperation Mail: cooperation Web: contact, cooperation	Coverage for income and age (Especially strong coverage bias for web!) Cumulative contact and cooperation for many other covariates Overall low Pseudo R^2	CAPI: Cooperation CATI: Coverage Mail: Cooperation Web: Coverage
Sequential Designs	Strong Improvement	Reduction of some bias for web, no improvement for mail Introduction of bias to web/mail where CAPI was biased	No improvement

Consistent with prior research, we found a hierarchy of response rates, where web was lowest and CAPI highest (Table 9). Also consistently, the non-contact rate was below ten percent for the interviewer modes. Mail, assessed by recall, did even better. Only web seemed to stand out in this respect, with a recall rate twice as high as that of subjects in the mail condition. This was surprising, as we expected that contact processes are similar in mail and web. We speculate that an envelope, which contains both a letter and a questionnaire, is more salient than an envelope containing only a letter. This perhaps contributes to establishing contact as well as its recall. An advice for practitioners therefore is to try to make the web invitation letter as visible as possible, e.g. by sending it in an envelope of comparable size to a letter containing also a paper questionnaire. However, non-cooperation rates were dominant in the determination of response rates, except for non-coverage in CATI, which was substantial as well. In particular in the web low cooperation leads to overall low response.

An assessment by relative (and absolute) sample bias changes the picture implied by response rates. No single mode was unbiased with respect to the population. In fact, nonresponse bias was strong indicated by many highly significant effects of socio-demographic indicators. Of all modes, it were the web and CAPI, which showed least significant indication for absolute bias, except for the income and ethnicity distribution (web) and age and ethnicity distributions (CAPI). It is remarkable that web showed bias on the fewest indicators, despite general worries about low response rates. We will come back to this finding below.

The interviewer modes (CAPI/CATI) and self-administered modes (web/mail) showed remarkable homogeneity. The only significant difference we found was on the gender distribution. Homogeneity still held in face of substantial coverage bias introduced in telephone and web. Generally, the typical socio-demographic patterns of non-coverage by telephone (younger, urban, single households, less non-Western foreigners) and Internet (lower income, old age, single households, less non-Western foreigners) were recognizable, as we expected. Curiously, however, this did not impact the homogeneity of CAPI/CATI and web/mail. The process of contact and cooperation balances the web and mail response to resemble each other more closely. CAPI tended to lose those respondents in the process of contact and cooperation lost due to coverage in CATI. Such mitigation is a new aspect in research on the impact of coverage, which often assumes coverage bias transcends to the final sample. We advise methodologists to consider this option in future research.

After we pooled the interviewer and non-interviewer modes to reduce complexity, we found more pronounced differences. Distinct coverage and differential contact and cooperation processes created mode-dependent sample bias mainly between these mode groups. The signs of effects of coverage on the one hand and cooperation and contact on the other hand, were mainly similar. These findings confirm our expectations about the differential processes of contact and cooperation. The interviewer is a strong reminding factor in the process of contact, but also a persuasive factor in the process of cooperation. However, his or her success varies clearly across the population. As these effects are absent in the non-interviewer modes, mode-dependent response behaviour is induced.

We did not find evidence for the presence of mode preferences beyond the interviewer-non-interviewer divide. For example, there were no pronounced differences in cooperation comparing web and mail or CAPI and CATI on the cooperation (and contact) stage. These results stress the relevance of the interviewer as major distinguishing – and biasing – characteristic between interviewer and self-administered modes.

We could not determine exactly, if differences between pooled non-interviewer and interviewer modes are due to differential contact or cooperation propensity, because the differences mainly appeared when considering the effects jointly (table 5, column 8). We interpret this as cumulative effect. That is, effect sizes are too small on both levels to be detected, but they are when analysed jointly. An alternative explanations, however, is measurement error in the recall based contact measure. We cannot finally conclude about the true origin of results on either the contact or cooperation stage.

The detailed analyses of bias on socio-demographics were supplemented by exploratory findings for the re-measurements of target variables from the Crime Victimisation Survey. We found bias on many target variables, but with one exception, bias was independent of mode. This also held for comparisons of pooled interviewer and non-interviewer modes indicating absence (or non-detectable) relative bias. These results confirm worries in literature about bias on a survey's target variables. However, they suggest independence of sample bias for the major modes of administration. This is surprising, as we generally would have expected to find at least some differences for target variables correlated with socio-demographic indicators. One alternative explanation for this finding is that the second wave data themselves might have been subject to nonresponse bias. Therefore we advise interpreting these results with care.

Finally, we presented assessments based on the mode-representativeness definition. It became clear that the conclusions about mode-dependency are not necessarily equivalent to the assessment based on relative sample bias. Assessing representativeness exploits response variance estimates as a summary measure. The R-indicator adds up all deviations from constant response variance in all sub-domains, also those non-significant effects suppressed when testing for significant relative sample bias. Moreover, if there is a strong deviation in only one domain, whilst constant variances in all others, this can still strongly influence the R-indicator. This situation was found in our study for the gender distribution in mail, which was the only significant difference between mail and web, for example, but still strongly influenced the R-indicator estimated for mail. Therefore, the R-indicator perspective does suggest less homogeneity between web and mail, despite being rather equivalent on the majority of indicators.

A virtue of the R-indicator is the quick overview it provides on changes of representativeness throughout the response process. It illustrates that the overall impact of non-coverage on representativeness is strong. Non-coverage presents the major loss of representativity in CATI and web. Moreover, we showed that representativeness evolves throughout the process of response. Surprisingly, web gained in representativeness due to its fortunate coverage and contact properties. Also CATI gained a bit. These processes are comparable to the ones leading to homogeneity of CAPI/CATI and mail/web found under the sample bias perspective. Moreover, note that CAPI and mail strongly loose in representativeness due to contact and cooperation, partly groups that already were lost due to non-coverage in CATI and web. In the outcome of the response process, representativeness differences between modes are therefore not strong.

However, it is remarkable that the web mode produced the most representative sample in the end, which reflects the small absolute sample bias and the related points made above (assessing absolute sample bias). We, therefore, argue that beliefs about the non-representativeness of web might be outdated by improved coverage in countries featuring high Internet coverage. However, processes of cooperation and contact were very beneficial. It, therefore, needs to be confirmed in other situations, whether the web mode performs equally well.

Though survey response is slightly mode-dependent, at least between interviewer and non-interviewer modes, sequential designs can only imperfectly make use of the effects. Consistent with expectations, sequential designs increased response rates drastically. For web, we found some bias reduction indicated by less or non-significant bias estimates. However, the strong sample bias of the income distribution of web could not be equalized. For the mail condition, we were even less successful in reducing bias. We conclude that mixing modes has the potential to reduce bias on variables, on which the first stage and second stage modes of a sequential design share opposite signed biases. However, for many variables CAPI or CATI do not possess the right response properties to balance the final sample. Tentatively, we may conclude that web followed by CAPI outperformed mail followed by CAPI. However, CAPI and CATI can also introduce bias for those variables that the modes are biased for but that web and mail are not biased for. We observed this, for example, for the urban agglomeration indicator. This is certainly a very undesirable side-effect of mixed-mode designs.

In sum, the rather homogenous response samples of the interviewer and non-interviewer modes suggest that only a mix of modes from these groups can usefully achieve superior samples, i.e. combining either CAPI and/or CATI with web and /or mail. However, even these mixed-mode designs do not fully work in practice, because the response properties of modes do not have sufficient complementarity. Mixed-mode produces a new sample, which might be biased for new variables and less biased for others. Accordingly, the representativeness, as assessed by the R-indicator, did not change in our case.

Survey researchers, however, still may consider mixed-mode designs useful or necessary. For example, the need for high response rates, to focus on particular sub-groups in later analysis (e.g. illiterate, migrants etc.), or to assure democratic principles in sampling (equal inclusion probability of all citizens) can all necessitate the use of more than one survey mode.

These results should be considered against some general remarks and limitations. Generally, the bias assessment depends on the socio-demographic or other auxiliary variables considered. If other auxiliary variables are used in the assessment, findings might differ. For example, the register of The Netherlands does not include information on educational attainment. Education, however, has been shown an important characteristic related to web coverage and general participation. We also do not know about bias on the survey's target variables.

Since our assessments quite consistently did only indicate small (socio-demographics) or no differences (target variables) between modes, we are optimistic to find similar results, if more variables were available for an assessment. In absence of strong differences on available variables, one can argue that reliance on the response rate becomes more important as indicators for selection effects that are 'hidden' from our assessment. Therefore a CAPI survey still may present the least threat of nonresponse bias as a single-mode design, and a mixed-mode design of web or mail followed by Capi or Capi/CATI might be superior to single-mode web or mail.

Furthermore, in assessing mode-dependency we advise to make a clear statement on the underlying response model and derived definition of mode-dependent (non-) response. The results of mode-dependency testing generally depend on statistical power of the design. In this study we believe sample sizes were sufficiently large to detect the stronger biases between modes. Noteworthy, our definitions and statistical modelling are based on conditional estimation. We included all main effects and interactions with the mode-indicators in the response models when testing sample bias. Earlier studies often used bivariate tests, for example for independence testing. In such situations it is likely to overestimate the true mode-dependency, since spurious relationships between response probabilities and auxiliary variables are not controlled. Hence, they may manifest in multiple bivariate associations of response and auxiliary variables. We, therefore, advise to rely on conditional models, if the overall extent of mode-dependent bias or representativeness is to be assessed. If the interest is bias of particular population domains, however, bivariate tests may be the more viable choice.

Appendix A: Assessing wave 2 independence and attrition

To generalize results from estimates based on the wave 2 CAPI/CATI survey, we assumed that the survey was a typical implementation of a CAPI/CATI survey. It could be argued, for example, that the first survey wave strongly affects some population domains in their response propensity to wave 2, regardless of survey topic, resulting in changes of response probability distributions. Indicators for sample quality include the response rate and nonresponse bias on background characteristics. In section 5, we already noted that overall response rates dropped due to reduced cooperation as compared to wave 1. This might not be problematic, if this drop was constant across relevant population domains. We go ahead testing for the CAPI mode in wave two, if C in

$$P(R_1 = 1 \mid M_1 = CAPI, M_2 = CAPI) = C * P(R_2 = 1 \mid M_1 = CAPI, M_2 = CAPI) \quad \text{(A1)}$$

is constant. Note that we compare first wave response probabilities to the CAPI mode with second wave probabilities in CAPI. First we test H0: $\beta = 0$ in the logistic random effect model:

$$\log it\left[P(R_{it} = 1 \mid M_1 = CAPI, M_2 = CAPI)\right] = \rho + \beta x_t + u_i \quad \text{(A2)}$$

Here, R_{it} is response (0,1) for individual i at either wave $t=1$ or $t=2$. $x_t = 0$ if $t=1$ and 1 otherwise. u_i is a Gaussian random effect over individuals with mean zero. We allow for random variation of initial response probabilities in the population and test whether the odds for response in wave two differ from those of wave 1. Estimating model (A2) gives $\hat{\beta} = -.828$ and $SE(\hat{\beta}) = .114$, which is significant. The odds for response in wave 2 are less than half of the odds of response in wave 1. This shows that response probabilities changed across waves.

We, now, go on by assessing whether the change in probabilities can be considered constant. One way of assessing this is hypothesizing β in model A2 to be invariant in any population characteristic X, i.e.

$$E(\beta_i \mid X) = \beta_0 \text{ and } Var(\beta_i \mid X) = 0$$

We can at least partially assess this assumption by testing H0: $\omega = 0$ in $E(\beta_i \mid X) = \beta_0 + X\omega$ by expanding (A2) for interactions between x_t and X_t:

$$\log it\left[P(R_{it} = 1 \mid M_1 = CAPI, M_2 = CAPI)\right] = \rho + (\beta_0 + X\omega)x_t + u_i \quad \text{(A3)}$$

In estimating this model in practice, we have register variables available as well as the survey mode a respondent was randomized to in wave 1 (cf. Table 1, socio-demographics). Without exception the hypothesis H0: $\omega = 0$ could not be rejected for any of the categories of X of wave 1 mode. We conclude that there was no systematic attrition in wave 2 as far as it can be assessed.

Appendix B: Detailed response models

Table B1: Response models of final response indicator ('outcome') by mode

	CAPI	CATI	Mail	Web	Interaction LRT		
					Overall	CAPI vs. CATI	Mail vs. Web
Gender (ref.: men)							
Female	-.156 (.104)	.357 (.104)***	.515 (.100)***	.047 (.108)	.0000	.0005	.0014
Age (ref.: 15-24)					n.s.	n.s.	n.s.
25-34	-.512 (.211)*	-.426 (.211)*	.054 (.193)	-.011 (.219)			
35-44	-.634 (.226)**	-.331 (.215)	.129 (.201)	.018 (.236)			
45-54	-.735 (.227)**	-.323 (.227)	.287 (.211)	.175 (.239)			
55-64	-.191 (.251)	.006 (.246)	.663 (.234)**	.475 (.258)			
65+	-.170 (.255)	.254 (.252)	.664 (.238)**	.103 (.268)			
Income (ref.: 1-15,000)					.0361	n.s.	n.s.
Up to 30,000	.324 (.137)*	.040 (.139)	-.003 (.126)	.226 (.149)			
Up to 45,000	.262 (.162)	.281 (.161)	.642 (.150)***	.817 (.164)***			
Up to 60,000	.295 (.204)	.333 (.211)	.798 (.207)***	.845 (.202)***			
60000+	.308 (.243)	.425 (.225)	.650 (.220)**	.717 (.236)**			
Missing	.144 (.174)	-.219 (.168)	-.109 (.159)	.033 (.178)			
Ethnicity (ref: Dutch)					.0077	n.s.	n.s.
Non-Western	-.803 (.167)***	-1.385 (.204)***	-.996 (.186)***	-.638 (.200)***			
Western	-.466 (.160)**	-.685 (.161)***	-.102 (.151)	-.046 (.167)			

(Continued on the next page)

	CAPI	CATI	Mail	Web	Interaction LRT: Overall	Interaction LRT: CAPI vs. CATI	Interaction LRT: Mail vs. Web
Marital Status (ref.: Single)					n.s.	n.s.	n.s.
Married, partnership	.254 (.152)	.318 (.154)*	.275 (.146)	.136 (.166)			
Divorced, widowed	.147 (.185)	-.200 (.195)	-.353 (.185)	-.393 (.210)			
HH size (ref.: 1 person)					n.s.	n.s.	n.s.
2 person	.288 (.156)	.494 (.167)**	.234 (.154)	.527 (.177)**			
3 person+	.465 (.169)**	.711 (.177)***	.069 (.159)	.444 (.189)*			
Degree of Urbanization (ref.:					n.s.	n.s.	n.s.
Strong	.069 (.154)	.300 (.163)	-.123 (.153)	-.062 (.167)			
Moderate	.150 (.175)	.225 (.181)	.077 (.172)	-.021 (.185)			
Little	.314 (.175)	.165 (.180)	-.095 (.171)	-.254 (.185)			
None	.627 (.214)**	.272 (.204)	-.077 (.197)	-.198 (.212)			
Urban agglomeration					n.s.	n.s.	n.s.
Amsterdam (ref.: other)	-.677 (.210)**	-.489 (.236)*	-.051 (.208)	-.086 (.232)			
Utrecht	.049 (.284)	-.272 (.311)	.407 (.296)	.262 (.307)			
Rotterdam	-.489 (.220)*	.005 (.231)	-.307 (.218)	-.342 (.247)			
Const.	.430 (.245)	-.858 (.258)***	-.853 (.237)***	-1.670 (.270)***			
N	2081	2062	2182	2199			
Pseudo R^2	.055	.079	.068	.052			
R-Indicator (adj.)	.765 (.021)	.697 (.019)	.717 (.019)	.797 (.020)			

Table B2: Response model of coverage and contact propensities by mode

	COVERAGE		CONTACT (cond. on coverage, recall of contact for Mail/Web)				Interaction LRT		
	CATI	web	CAPI	CATI	Mail	Web	Overall	CAPI vs. CATI	Mail vs. Web
Gender female (ref.: men)	.337 (.113)**	.050 (.139)	.196 (.212)	-.021 (.198)	.412 (.208)*	-.186 (.181)	n.s.	n.s.	.0301
Age (ref.: 15-24)							.0003	n.s.	n.s.
25-34	-.389 (.218)	-1.120 (.624)	-.682 (.396)	-.819 (.382)*	.155 (.410)	.217 (.338)			
35-44	.001 (.229)	-2.130 (.590)***	-.817 (.436)	-.724 (.385)	-.162 (.421)	.588 (.380)			
45-54	-.035 (.242)	-2.317 (.595)***	-.200 (.455)	-.677 (.424)	-.264 (.457)	.617 (.388)			
55-64	.431 (.269)	-3.056 (.599)***	.471 (.521)	-.325 (.474)	-.123 (.524)	.904 (.451)*			
65+	.904 (.277)***	-4.873 (.598)***	1.877 (.651)**	.484 (.530)	-.493 (.524)	.805 (.482)			
Income (ref.: 1-15,000)							n.s.	n.s.	n.s.
Up to 30,000	.099 (.154)	.590 (.154)***	.222 (.311)	-.213 (.295)	.300 (.269)	.071)(.248)			
Up to 45,000	.189 (.178)	1.656 (.235)***	.315 (.347)	-.103 (.331)	.431 (.316)	.457 (.296)			
Up to 60,000	.628 (.254)*	2.849 (.537)***	-.271 (.379)	-.257 (.406)	1.156 (.522)*	.141 (.345)			
60000+	.511 (.264)	1.961 (.431)***	-.468 (.461)	-.141 (.453)	.697 (.473)	-.206 (.382)			
Missing	.021 (.184)	.744 (.300)*	.135 (.389)	-.408 (.330)	.373 (.354)	-.169 (.271)			
Ethnicity (ref: Dutch)							n.s.	n.s.	n.s.
Non-Western	-1.234 (.175)***	-1.323 (.281)***	-.591 (.283)*	-.394 (.370)	-.570 (.333)	-.287 (.293)			
Western	-.485 (.165)**	.095 (.215)	-.766 (.291)**	-.418 (.296)	-.614 (.291)*	.182 (.336)			
Marital Status (ref.: Single)							n.s.	n.s.	n.s.
Married, partnership	.095 (.165)	.463 (.273)	.412 (.307)	1.009 (.271)***	.494 (.333)	-.073 (.294)			
Divorced, widowed	-.182 (.206)	-.192 (.241)	-.264 (.350)	.612 (.369)	-.376 (.379)	-.578 (.372)			
HH size (ref.: 1 person)							.0001	n.s.	n.s.
2 person	.517 (.171)**	1.333 (.226)***	1.357 (.301)***	.390 (.293)	.392 (.350)	.113 (.337)			
3 person+	.784 (.179)***	1.332 (.266)***	1.136 (.286)***	.481 (.306)	-.579 (.335)	-.217 (.339)			
Urbanity (ref.: Very strong)							.0099	n.s.	.0075
Strong	.349 (.166)*	-.047 (.224)	.101 (.280)	.382 (.326)	.715 (.318)*	-.792 (.340)*			
Moderate	.434 (.191)*	.081 (.249)	.793 (.392)*	-.032 (.345)	.657 (.351)	-.708 (.362)*			
Little	.739 (.197)***	-.353 (.243)	.717 (.379)	.157 (.348)	.500 (.339)	-.728 (.362)*			
None	.641 (.228)**	-.462 (.270)	1.222 (.575)*	.268 (.402)	.977 (.451)*	-.974 (.389)*			
Urban agglomeration							n.s.	.0482	n.s
Amsterdam (ref.: other)	-.262 (.227)	.124 (.348)	-1.021 (.319)**	.117 (.485)	-.062 (.438)	-.494 (.455)			
Utrecht	-.111 (.311)	.514 (.481)	.750 (.647)	-.260 (.520)	.561 (.598)	-.163 (.610)			
Rotterdam	-.222 (.234)	-.101 (.309)	-.399 (.386)	.439 (.531)	.138 (.455)	.079 (.588)			
Const.	-.277 (.262)	3.893 (.588)***	1.701 (.434)***	1.813 (.491)***	1.274 (.467)**	1.808 (.489)**			
N	2062	4296	2081	1474	1112	972			
Pseudo R^2	.109	.366	.183	.081	.067	.041			
R-Indicator (adj.)	.754 (.028)	.664 (.012)	.857 (.018)	.893 (.021)	.923 (.019)	.901 (.030)			

Table B3: Response model of cooperation propensities by mode

	COOPERATION (conditional on contact)				Interaction LRT		
	CAPI	CATI	Mail	Web	Overall	CAPI vs. CATI	Mail vs. Web
Gender female (ref.: men)	-.196 (.109)	.365 (.146)*	.463 (.169)**	-.021 (.167)	.0032	.0020	.0415
Age (ref.: 15-24)					.0499	n.s.	n.s.
25-34	-.443 (.225)*	.198 (.335)	.033 (.313)	.052 (.334)			
35-44	-.549 (.241)*	-.191 (.316)	.325 (.341)	.058 (.359)			
45-54	-.774 (.239)**	-.205 (.335)	.515 (.365)	.214 (.371)			
55-64	-.287 (.264)	-.160 (.363)	1.045 (.409)*	.678 (.412)			
65+	-.361 (.267)	-.233 (.372)	1.024 (.419)*	.820 (.433)			
Income (ref.: 1-15,000)					n.s.	n.s.	n.s.
Up to 30,000	.281 (.143)	.045 (.189)	.246 (.222)	.079 (.229)			
Up to 45,000	.211 (.169)	.387 (.231)	.867 (.268)**	.633 (.251)*			
Up to 60,000	.387 (.221)	.166 (.287)	.433 (.327)	.490 (.307)			
60000+	.422 (.264)	.361 (.312)	.659 (.373)	.372 (.362)			
Missing	.127 (.181)	-.351 (.229)	.042 (.275)	.207 (.276)			
Ethnicity (ref: Dutch)					n.s.	n.s.	n.s.
Non-Western	-.783 (.176)***	-1.061 (.299)***	-1.013 (.303)**	-.685 (.307)*			
Western	-.361 (.171)*	-.633 (.215)**	.195 (.293)	-.082 (.274)			
Marital Status (ref.: Single)					n.s.	n.s.	n.s.
Married, partnership	.248 (.159)	.020 (.234)	.208 (.255)	.137 (.259)			
Divorced, widowed	.136 (.197)	-.625 (.278)*	.049 (.339)	-.286 (.338)			
HH size (ref.: 1 person)					n.s.	n.s.	n.s.
2 person	.052 (.168)	.264 (.243)	.252 (.270)	.202 (.299)			
3 person+	.227 (.183)	.317 (.269)	.216 (.275)	.314 (.311)			
Urbanity (ref.: Very strong)					n.s.	.0062	n.s.
Strong	.063 (.162)	-.003 (.243)	-.224 (.278)	.292 (.279)			
Moderate	.060 (.183)	-.094 (.271)	-.115 (.307)	.208 (.299)			
Little	.235 (.183)	-.518 (.262)*	-.154 (.306)	-.120 (.296)			
None	.522 (.223)*	-.284 (.293)	-.209 (.349)	.190 (.335)			
Urban agglomeration					n.s.	n.s.	n.s.
Amsterdam (ref.: other)	-.526 (.225)*	-.646 (.315)*	.129 (.399)	.584 (.442)			
Utrecht	-.064 (.295)	-.295 (.462)	.040 (.480)	1.222 (.591)*			
Rotterdam	-.463 (.230)*	.144 (.367)	-.228 (.403)	.275 (.455)			
Const.	.875 (.262)***	1.118 (.395)**	-.073 (.393)	-.681 (.433)			
N	1963	1338	983	777			
Pseudo R^2	.039	.044	.078	.050			
R-Indicator (adj.)	.822 (.022)	.840 (.026)		.811 (.036)			

References

Adua, L., & Sharp, J. S. (2010). Examining survey participation and response quality: The significance of topic salience and incentives. *Survey Methodology*, *36*(1), 95–109.

Agresti, A. (2002). *Categorical Data Analysis*. Hoboken, New Jersey: Wiley-Interscience.

Bälter, K. A., Bälter, O., Fondell, E., & Lagerros, Y. T. (2005). Web-based and Mailed Questionnaires. *Epidemiology*, *16*(4), 577–579.

Berrens, R. P., Bohara, A. K., Jenkins Smith, H., Silva, C., & Weimer, D. L. (2003). The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples. *Political Analysis*, *11*(1), 1 –22.

Bethlehem, J. (1988). Reduction of Nonresponse Bias Through Regression Estimation. *Journal of Official Statistics*, *4*(3), 251–260.

Bethlehem, J. (2002). Weighting Nonresponse Adjustment Based on Auxiliary Information. In R. M. Groves, D. A. Dillman, J. Eltinge, & R. J. A. Little (Eds.), *Survey Nonresponse*, Wiley Series in Probability and Statistics. New York: Wiley & Sons.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to Survey Quality*. New Jersey: Wiley.

Billiet, J., Philippens, M., Fitzgerald, R., & Stoop, I. (2007). Estimation of nonresponse bias in the European Social Survey: using information from reluctant respondents. *Journal of official statistics*, *23*(2), 135–162.

Blumberg, S. J., Luke, J. V., Cynamon, M. L., & Frankel, M. R. (2008). Recent Trends in Household Telephone Coverage in the United States. In J. Lepkowski, C. Tucker, J. M. Brick, E. D. De Leeuw, L. Japec, P. J. Lavrakas, M. W. Link, et al. (Eds.), *Advances in Telephone Survey Methodology* (pp. 56–86). New York: Wiley.

Blyth, B. (2008). Mixed mode: the only “fitness” regime. *International Journal of Market Research*, *50*(2), 241–266.

Buelens, B., Laan, J. van der, Schouten, B., Brakel, J. van den, Burger, J., Klausch, T. (2012), Disentangling mode-specific selection and measurement bias in social surveys, Discussion paper 201211, Statistics Netherlands.

Busse, B., & Fuchs, M. (2009). Relative coverage biases for mobile phone and mobile web surveys. In N. Döring, A. Liess, & E. Maxl (Eds.), *Mobile Market Research. Neue Schriften zur Online-Forschung Band 2* (pp. 217–236). Köln: Herbert von Halem.

Busse, B., & Fuchs, M. (2011). The components of landline telephone survey coverage bias. The relative importance of no-phone and mobile-only populations. *Quality & Quantity*.

Chang, L., & Krosnick, J. A. (2009). National Surveys Via Rdd Telephone Interviewing Versus the Internet. *Public Opinion Quarterly*, *73*(4), 641–678.

Converse, P. D., Wolfe, E. W., Xiaoting Huang, & Oswald, F. L. (2008). Response Rates for Mixed-Mode Surveys Using Mail and E-mail/Web. *American Journal of Evaluation*, *29*(1), 99 –107.

Couper, M. P. (2000). Review: Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly*, *64*(4), 464–494.

Couper, M. P., Kapteyn, A., Schonlau, M., & Winter, J. (2007). Noncoverage and nonresponse in an Internet survey. *Social Science Research*, *36*(1), 131–148.

Cuppen, M., Laan, P. van der, Nunspeet, W. van (2011), Re-engineering Dutch social surveys: From single-purpose surveys to an integrated design, Paper presented at the ISI World Statistics Congress, Dublin, Ireland, 21-26 August.

De Heer, W. (1999). International response trends: results of an international survey. *Journal of Official Statistics*, *15*(2), 129–142.

De Leeuw, E. (1992). *Data Quality in Mail, Telephone, and Face to Face surveys*. Amsterdam: TT-Publicaties.

De Leeuw, E. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, *21*(2), 233–255.

De Leeuw, E., & De Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In Robert M. Groves, D. A. Dillman, J. Eltinge, & R. Little (Eds.), *Survey Nonresponse* (pp. 41–54). New York: Wiley.

De Leeuw, E., Dillman, D. A., & Hox, J. (2008). Mixed mode surveys: When and Why. *International Handbook of Survey Methodology*. New York: Lawrence Erlbaum.

De Leeuw, E., & Hox, J. (2008). Self-administered questionnaires: mail surveys and other applications. In E. De Leeuw, J. J. Hox, & D. A. Dilman (Eds.), *International Handbook of Survey Methodology* (pp. 239–263). New York: Lawrence Erlbaum.

De Leeuw, E., & Hox, J. J. (2010). Internet Surveys as Part of a Mixed-Mode Design. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet: advances in applied methods and research strategies* (pp. 45–76). Routledge.

Dijkstra, W., & Smit, J. H. (2002). Persuading Reluctant Recipients in Telephone Surveys. In R. Groves, D. Dillman, J. Eltinge, & R. Little (Eds.), *Survey Nonresponse*. New: Wiles & Sons.

Dillman, D. A., & Christian, L. M. (2005). Survey Mode as a Source of Instability in Responses across Surveys. *Field Methods*, *17*(1), 30 –52.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research*, *38*(1), 1–18.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, Mail, and Mixed-Mode Surveys. The Tailored Design Method.* New Jersey: Wiles & Sons.

Dillman, D., West, K. K., & Clark, J. R. (1994). Influence of an Invitation to Answer by Telephone on Response to Census Questionnaires. *Public Opinion Quarterly*, *58*(4), 557–568.

Diment, K., & Garrett-Jones, S. (2007). How Demographic Characteristics Affect Mode Preference in a Postal/Web Mixed-Mode Survey of Australian Researchers. *Social Science Computer Review*, *25*(3), 410 –417.

Durrant, G. B., & Steele, F. (2009). Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*, 361–381.

Ellis, C. H., & Krosnick, J. A. (1999, April). *Comparing Telephone and Face-to-Face Surveys in Terms of Sample Representativeness: A Meta-Analysis of Demographic Characteristics*. Ohio State University, Ohio, USA. Retrieved from ftp://ftp.nes.isr.umich.edu/ftp/resourcs/techrpts/reports/Tch59.pdf

Fowler, F. J., Gallagher, P. M., Stringfellow, V. L., Zaslavsky, A. M., Thompson, J. W., & Cleary, P. D. (2002). Using Telephone Interviews to Reduce Nonresponse Bias to Mail Surveys of Health Plan Members. *Medical Care*, *40*(3), 190–200.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An Experimental Comparison of Web and Telephone Surveys. *Public Opinon Quarterly*, *69*(3), 370–392.

Fuchs, M. (2008). Mobile web survey: a preliminary discussion of methodological implications. In F. G. Conrad & M. F. Schober (Eds.), *Envisioning the Survey Interview of the Future* (pp. 77–94). New York: Wiley.

Greene, J., Speizer, H., & Wiitala, W. (2008). Telephone and Web: Mixed-mode Challenge. *Health Services Research*, 43, 230-248 .

Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, *56*(4), 475–495.

Groves, Robert M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.

Groves, Robert M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, *70*(5), 646–675.

Groves, Robert M., & Couper, M. (1998). *Nonresponse in household interview surveys*. New Jersey: Wiley.

Groves, Robert M., & Couper, M. P. (1996). Contact-Level Influences on Cooperation in Face-to-Face Surveys. *Journal of Official*, *12*(1), 63–83.

Groves, Robert M., Couper, M. P., Presser, S., Singer, E., Tourangeau, R., Acosta, G. P., & Nelson, L. (2006). Experiments in Producing Nonresponse Bias. *Public Opinion Quarterly*, *70*(5), 720 –736.

Groves, Robert M., & Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias. *Public Opinion Quarterly*, *72*(2), 167 –189.

Groves, Robert M., Presser, S., & Dipko, S. (2004). The Role of Topic Interest in Survey Participation Decisions. *Public Opinion Quarterly*, *68*(1), 2 –31.

Groves, Robert M., Singer, E., & Corning, A. (2000). Leverage-Saliency Theory of Survey Participation: Description and an Illustration. *Public Opinion Quarterly*, *64*(3), 299–308.

Heerwegh, D., & Loosveldt, G. (2011). Assessing Mode Effects in a National Crime Victimisation Survey using Structural Equation Models: Social Desirability Bias and Acquiescence. *Journal of Official Statistics*, *27*(1), 49–63.

Hox, J. J., & Leeuw, E. D. (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys. *Quality & Quantity*, *28*(4), 329–344.

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the Effect of Data Collection Mode on Measurement. *International Statistical Review*, *78*(1), 3–20.

Kaplowitz, M. D., Hadlock, T. D., & Levine, R. (2004). A Comparison of Web and Mail Survey Response Rates. *Public Opinion Quarterly*, *68*(1), 94–101.

Keeter, S., Miller, C., Kohut, A., Groves, R., & Presser, S. (2000). Consequences of Reducing Nonresponse in a National Telephone Survey. *Public Opinion Quarterly*, *64*(2), 125 –148.

Kwak, N., & Radler, B. T. (2002). A comparison between mail and Web surveys. Response pattern, respondent profile, and data quality. *Journal of Official Statistics*, *18*(2), 257–74.

Link, M. W., & Mokdad, A. H. (2005). Alternative Modes for Health Surveillance Surveys: An Experiment with Web, Mail, and Telephone. *Epidemiology*, *16*(5), 701–704.

Link, M. W., & Mokdad, A. H. (2006). Can Web and Mail Survey Modes Improve Participation in an RDD-based National Health Surveillance? *Journal of Official Statistics*, *22*(2), 293–312.

Lynn, P., & Clarke, P. (2002). Separating Refusal Bias and Non-Contact Bias: Evidence from UK National Surveys. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *51*(3), 319–333.

Lynn, P., Uhrig, S. C. N., & Burton, J. (2010). *Lessons from a Randomised Experiment with Mixed-Mode Designs for a Household Panel Survey*. Understanding Society Working Paper Series, No. 2010-03, Essex. Retrieved from http://www.iser.essex.ac.uk/publications/working-papers/understanding-society/2010-03

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes A meta-analysis comparing response rates. *International Journal of Market Research*, *50*(1), 79–104.

Manfreda, K. L., & Vehovar, V. (2008). Internet Surveys. In E. De Leeuw, J. J. Hox, & D. A. Dilman (Eds.), *International Handbook of Survey Methodology* (pp. 264–284). New York: Lawrence Erlbaum.

Miller, T. I., Kobayashi, M. M., Caldwell, E., Thurston, S., & Collett, B. (2002). Citizen Surveys on the Web. *Social Science Computer Review*, *20*(2), 124 –136.

Mohorko, A. (2011). *Surveying the General Population: Coverage bias in European Web and Telephone Surveys* (Unpublished MA-Thesis). Utrecht University, Department of Methodology and Statistics, Utrecht, The Netherlands.

Morton-Williams, J. (1993). *Interviewer Approaches*. Aldershot: Dartmouth.

Nicoletti, C., & Peracchi, F. (2005). Survey response and survey characteristics: microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(4), 763–781.

Porter, S. R., & Whitcomb, M. E. (2007). Mixed-Mode Contacts In Web Surveys. *Public Opinion Quarterly*, *71*(4), 635 –648.

Särndal, C.-E., & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Chichester: Wiley.

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing Response Rates and Nonresponse Bias in Web and Paper Surveys. *Research in Higher Education*, *44*(4), 409–432.

Schouten, B. (2012), Mixed-mode survey designs. Recent developments at Statistics Netherlands, Paper presented at the 4th International Workshop on Internet Survey Methods (IWIS), 12 and 13 September, Daejeon, Korea.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, *35*(1), 101–113.

Schouten, B., Shlomo, N., & Skinner, C. (2011). Indicators for Monitoring and Improving Representativeness of Response. *Journal of Official Statistics*, *27*(2), 231–253.

Shih, T.-H., & Xitao Fan. (2008). Comparing Response Rates from Web and Mail Surveys: A Meta-Analysis. *Field Methods*, *20*(3), 249 –271.

Shlomo, N., Skinner, C., & Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response. *Journal of Statistical Planning and Inference*, (142), 201–211.

Snijkers, G., Hox, J., & de Leeuw, E. (1999). Interviewers' Tactics for Fighting Survey Nonresponse. *Journal of Official Statistics, 15*(2), 185–198.

Thornberry, O. T., & Massey, J. T. (1988). Trends in United States Telephone Coverage Across Time and Subgroups. In Robert M. Groves, P. P. Biemer, & L. E. Lyberg (Eds.), *Telephone Survey Methodology* (pp. 25–50). John Wiley and Sons.

Trewin, D., & Lee, G. (1988). International Comparisons of Telephone Coverage. In Robert M. Groves, P. P. Biemer, & L. E. Lyberg (Eds.), *Telephone Survey Methodology* (pp. 9–24). John Wiley and Sons.

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A Method for Evaluating Mode Effects in Mixed-mode Surveys. *Public Opinion Quarterly*, *74*(5), 1027 –1045.

Voogt, R. (2004). *"I'm Not Interested". Nonresponse bias, response bias and stimulus effects in election research.* University of Amsterdam, Amsterdam.