DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 1, 2015, pp. 1–30, http://dx.doi.org/10.1515/JOS-2015-0001

Face-to-Face or Sequential Mixed-Mode Surveys Among Non-Western Minorities in the Netherlands: The Effect of Different Survey Designs on the Possibility of Nonresponse Bias

Johannes W.S. Kappelhof[1]

This article compares the quality of response samples based on a single mode CAPI survey design with the quality of response samples based on a sequential mixed-mode (CAWI-CATI-CAPI) survey design among four non-Western minority ethnic groups in the Netherlands. The quality is assessed with respect to the representativity of the response samples and the estimated potential for nonresponse bias in survey estimates based on auxiliary variables and the response rate. This article also investigates if these designs systematically enhance response rates differently among various sociodemographic subgroups based on auxiliary variables. Also, costs and cost-related issues particular to this sequential mixed-mode design are discussed. The results show that sequential mixed mode surveys among non-Western ethnic minorities in the Netherlands lead to less representative response samples and show more potential for nonresponse bias in survey estimates. Furthermore, the designs lead to systematic differences in response rates among various sociodemographic subgroups, such as older age groups. Both designs also cause some of the same sociodemographic subgroups to be systematically underrepresented among all non-Western ethnic minority groups. Finally, the results show that in this instance the cost savings did not outweigh the reduction in quality.

Key words: Survey design; sequential mixed-mode survey; nonresponse bias; non-western ethnic minorities; representativeness.

1. Introduction

In general population surveys, minority ethnic groups tend to be underrepresented (Feskens 2009; Groves and Couper 1998; Schmeets 2005; Stoop 2005). At the same time, national and international policy makers need specific information about these groups, especially on issues such as socioeconomic and cultural integration (Bijl and Verweij 2012). That is why separate surveys among the main minority ethnic groups, that is non-Western minorities, continue to be necessary in the Netherlands. However, large-scale surveys are costly, and surveys among minorities are even more expensive per completed interview than general surveys, due to the lower response rates among minorities. It is therefore of great importance to determine which strategies are effective for surveying ethnic minorities, while maintaining an acceptable level of quality and minimizing the costs.

One important part of the survey design is the data-collection mode (face-to-face, telephone, web or paper). These modes vary greatly not only in costs, but also in the

[1] The Netherlands institute for Social Research/SCP, The Hague, P.O. Box 16164, The Netherlands. Email: j.kappelhof@scp.nl
Acknowledgment: I would like to thank E.D. de Leeuw, I.A.L. Stoop and some anonymous reviewers for their very helpful comments on previous drafts of this article.

probability of completing an interview, especially among non-Western minorities (Feskens et al. 2010). There are reasons to believe that these groups may not be as well represented if a survey is conducted by means of less expensive data-collection modes as compared to a single-mode face-to-face survey. Telephone, web and mail questionnaires all lead to increased nonresponse due to higher refusal rates, a higher prevalence of functional illiteracy and/or lower penetration rates of modes compared to face-to-face (Dagevos and Schellingerhout 2003; Feskens 2009; Feskens et al. 2010; Gijsberts and Iedema 2011; Kappelhof 2010; Kemper 1998; Korte and Dagevos 2011; Schmeets 2005; Schothorst 2002; Van Ingen et al. 2007; Veenman 2002).

Despite the known limitations of other modes of data collection, there is a strong push to explore the possibility of employing less expensive methods of data collection among non-Western minorities. One possible way of reducing costs and dealing with the additional nonresponse brought about by the different modes is through the use of a sequential mixed-mode survey (De Leeuw 2005).

This article sets out to investigate:

1. how the use of a sequential mixed-mode design in surveys among non-Western minorities in the Netherlands affects the quality of the *response* sample (i.e., the composition of the group of respondents) compared to a single-mode face-to-face design, and how these two designs can potentially impact nonresponse bias. This will be referred to as the *overall quality* research question.
2. whether these designs systematically enhance response rates differently among various socio-demographic subgroups among non-Western minorities. This will be referred to as the *systematic differences* research question.
3. Finally, we will discuss costs and cost-related issues particular to this sequential mixed-mode design that are relevant in the quality versus costs trade-off decision.

The data used in this study come from a large-scale survey design experiment. Two random samples were drawn from each of the four largest non-Western minority populations living in the Netherlands. Subsequently, one sample was assigned to a face-to-face computer-assisted personal interviewing (CAPI) design and the other sample was assigned to a sequential mixed-mode design using computer-assisted web interviewing (WEB), computer-assisted telephone interviewing (CATI) and face-to-face CAPI. The fieldwork for both survey conditions was conducted simultaneously by Gfk Netherlands and lasted from November 2010 until June 2011.

In this article, we are analyzing exclusively the representativity of the *response* samples and the estimated potential for nonresponse bias based on auxiliary variables and the response rate. However, we shall not compare actual estimates of substantive variables from both survey designs as an indication of the nonresponse bias related to the estimates, given that, in this experimental design, observed differences can also be (partly) caused by mode effects in the sequential mixed-mode design (De Leeuw 2005; De Leeuw et al. 2008; Dillman and Christian 2005; Voogt and Saris 2005). Furthermore, sampling error can also contribute to observed differences, although this can be estimated.

The article presents a brief overview of the main difficulties in data collection resulting in nonresponse when surveying non-Western minorities and how survey design can reduce these difficulties. The data and methods section describes the experiment in more detail and the methods used to answer our research aims. This is followed by the results of the analysis and the subsequent conclusion and discussion.

2. The Underrepresentation of Non-Western Minorities in Population Surveys in the Netherlands and Survey Design Choices

Statistics Netherlands uses the following official definition to describe a non-Western person in the Netherlands: "Every person residing in the Netherlands of whom one or both parents were born in Africa, Latin America, Asia (excluding Indonesia and Japan) or Turkey (Reep 2003)". A further distinction is made between first generation (born in Africa, Latin America and Asia (excluding Indonesia and Japan) or Turkey and moved to the Netherlands) and second generation (born in the Netherlands, but one or both parents were born in Africa, Latin America and Asia – excluding Indonesia and Japan – or Turkey). Indonesian and Japanese immigrants are seen as (more similar to) Western minorities based on their socioeconomic and sociocultural position, which mainly involves persons born in the former Dutch East Indies (Indonesia) and employees working for Japanese companies with their families. In 2011, non-Western minorities made up about 11% of the population in the Netherlands (CBS-Statline).

The main reason for the underrepresentation of non-Western minorities in population surveys in the Netherlands is nonresponse. A distinction can be made between direct causes and correlates for nonresponse. For instance, a direct cause would be language problems or the higher rate of illiteracy, especially among older non-Western immigrants (Feskens et al. 2010). A correlate would be that non-Western minorities more often tend to live in the larger cities in the Netherlands. Big-city dwellers in general are more difficult to contact and refuse more often (Groves and Couper 1998; Stoop 2005).

Adapting the survey design in such a way that these direct causes of nonresponse are addressed may reduce the nonresponse among non-Western minorities. Language difficulties stop being an issue if the design includes a translated questionnaire. Functional illiteracy ceases to be a problem when the interviews are conducted by interviewers who read out the questionnaire. Moreover, the use of the telephone for interviews increases the number of refusals among non-Western minorities to an incomparable degree as opposed to native Dutch or to a face-to-face mode and should therefore be avoided (Schothorst 2002).

Other cultural differences influencing nonresponse may also be reduced by specific survey design choices. For example, the use of interviewers with a common ethnic background: not only do they speak the language, but they are also aware of the proper etiquette for approaching the sampled persons. An often overlooked cause of nonresponse is the timing and length of the fieldwork. Especially among some of the ethnic minority groups, it is not uncommon to go on an extended holiday to their country of origin during the summer. Sometimes there is also a mismatch between religious holidays of ethnic groups and the way the agency plans the fieldwork (Kemper 1998; Schothorst 2002; Veenman 2002).

Sampling frame errors and especially undercoverage provide another reason why non-Western minorities are underrepresented in population surveys in the Netherlands. Undercoverage occurs when not all elements of the target population can be found in the sampling frame (Groves 1989). In the Netherlands, (semi)-governmental and scientific institutes mainly use the postal data service (delivery sequence file) or population register as a sampling frame. Both frames suffer from frame errors, such as mobility of the sample units, no known address of the sample units, slow registration of the sample units or death of the sample units. Some of these causes occur far more often among non-Western minorities, such as mobility or no known address of sample units (Feskens 2009; Kappelhof 2010).

3. Data and Methods

3.1. Data

The Dutch Survey on the Integration of Minorities (SIM) sets out to measure the socioeconomic position of non-Western minorities as well as their sociocultural integration. This survey is a nationwide, cross-sectional survey conducted every four years starting in 2006. A large-scale survey design experiment was conducted in the 2010–2011 SIM round.

In total, Statistics Netherlands drew ten samples: two random samples of named individuals were drawn from each of five mutually exclusive population strata; Dutch of Turkish, Moroccan, Surinamese, and Antillean (including Aruba) descent and the remainder of the population (mostly native Dutch) living in the Netherlands, aged 15 years and above. The present study focuses on how different designs affect the quality of the *response* sample and how they can potentially impact nonresponse bias in surveys conducted among non-Western minorities in the Netherlands. This is why the samples containing native Dutch are excluded from this article. The analysis is therefore based on eight samples.

Based on the official definition of non-Western minorities we will use a more narrow definition to define Dutch of Turkish, Moroccan, Surinamese, and Antillean descent to include persons that were either born in Turkey, Morocco, Surinam or the Dutch Antilles or have at least one parent who was born there. In cases where the father and mother were born in different countries, the mother's country of birth is dominant, unless the mother was born in the Netherlands, in which case the father's country of birth is dominant. These four ethnic groups make up about two thirds of the total non-Western population in the Netherlands (CBS-Statline). For the purpose of brevity, they will be referred to as Turkish, Moroccans, Surinamese and Antilleans in the remainder of this article.

From each ethnic group, one sample was allocated to a single-mode face-to-face CAPI design (SM) and one sample was allocated to a sequential mixed-mode design (MM). In the SM design, a minimum of three face-to-face contact attempts had to be conducted. The SM also included a limited reissue in which unsuccessful addresses were reissued to another CAPI interviewer who had to conduct another minimum of three face-to-face contact attempts.

In the MM design, all sample units were first sent an invitation to participate via WEB. Up to two reminders were sent to nonresponding sample units. Subsequently the remaining nonrespondents with a known fixed phone number were approached using CATI. Nonrespondents were called on at least four different days in the week, at different time periods during the day. If there was no answer or a busy signal, the number would be called more than once within the same time period. Finally, both the WEB-nonrespondents without a known (fixed) phone number and the CATI nonrespondents were approached using face-to-face interviewers (CAPI). WEB and CATI nonresponders were contacted at least three times by a face-to-face interviewer on different days and at different time periods. CATI was added as a mode, despite previous research indicating that this was not an optimal mode for surveying ethnic minorities. This was done in order to see whether this result was still valid a decade later, especially since the second-generation immigrants are much more familiar with telephones nowadays, but mostly to see if the use of CATI could potentially lead to cost savings.

In both survey designs standard response-enhancing measures were applied, such as advance letters, incentives and the possibility for potential respondents to call a toll-free number in case of questions or in order to reschedule an appointment for an interview.

This experiment used the population register as a sampling frame and the same stratified two-stage probability sampling design in all four population strata to draw the samples. In the first stage municipalities were selected proportional to size and in the second stage a fixed number of named individuals were selected. The strata variable used was municipality size and consisted of three strata: the four largest municipalities, all with a population of over 250,000; midsize municipalities with a population of between 50,000 and 250,000; and small municipalities with a population of less than 50,000. For each target group, the sample size was proportionally allocated across different municipality size strata (Table 1).

Process data and auxiliary information, also known as paradata, are potentially useful for increasing participation, for nonresponse adjustment or for evaluating potential nonresponse bias in survey estimates (Couper 2005; Kreuter 2013; Maitland et al. 2009). In this study we use the SIM fieldwork data files. These contain both process data, such as number, time, date, and outcome of contact attempt, and auxiliary information from the sampling frame about each sample unit, such as ethnicity, age, gender, first- or second-generation immigrants, municipality, and so on.

Differences Between Survey Designs

Besides the differences in administered mode and the use of a reissue phase, there is another important aspect that varied between both survey designs that could influence the results. The average length of the questionnaire differed between modes. The estimated average length of the questionnaire in the CAPI mode, based on CAPI timers, was about

Table 1. Gross sample sizes per ethnic group and design across municipality strata

	Turkish		Moroccans		Surinamese		Antilleans	
	SM	MM	SM	MM	SM	MM	SM	MM
Large municipalities	554	344	812	502	1020	633	695	429
Midsize municipalities	727	459	674	422	662	424	945	594
Small municipalities	284	176	254	162	248	150	334	210
Total	1,565	979	1,740	1,086	1,930	1,207	1,974	1,233

45 minutes. A 45-minute questionnaire was considered too long for both CATI and WEB by fieldwork experts and experts on minority research (Feskens et al. 2010). As a result, the questionnaire length for WEB and CATI has been reduced to an estimated 30 minutes.

Another difference between the designs is the value of the conditional or promised nonmonetary incentive. The use of incentives has a proven positive effect on response rates (Dillman 2007; Groves and Couper 1998; Singer et al. 1999; Singer et al. 2000; Singer 2002). In both designs a gift certificate was used as a promised incentive. In the SM design these gift certificates were worth €10. In the MM design the amount varied: €7.50 in the WEB mode and €10 in the other modes. As mentioned above, a maximum of two reminders was sent during the WEB phase to nonresponding sampled persons. After the second reminder the worth of the conditional non-monetary incentive was increased to €12.50. As both designs used conditional incentives and the difference in value was rather small, we believe this difference between survey conditions to have a minor impact on the results.

Differences in Survey Design Between Ethnic Groups

A recent survey conducted by Statistics Netherlands among the four largest non-Western minorities discovered that approximately 14% of the sample were nonrespondents due to language problems (Feskens 2009). Results from other surveys among the same minorities groups in the Netherlands showed that nonrespondents who are not able to read or speak Dutch are found mostly among the Turkish and Moroccan populations (Kappelhof 2010). For the SIM survey, auxiliary information about ethnicity, age, gender, municipality, and status as first- or second- generation immigrants was available in the sample frame data for all sampled persons. This allowed for a tailored approach for the sampled persons. Two types of tailoring were used in both arms of the experiment to increase response. They mainly have to do with anticipated language difficulties, but also with anticipated cultural differences. Research has shown that a greater cultural familiarity due to a shared ethnic background of interviewer and respondent may also be a factor in increasing the willingness to respond (see for instance Moorman et al. 1999).

The first type of tailoring was the use of translated questionnaires and advance letters. These were used in both designs in all modes (WEB, CATI, and CAPI), but only among the Moroccan and Turkish samples. Furthermore, a phonetically translated Berber version was available as an aid for the interviewer. This is a spoken (i.e., not written) language that many Moroccans living in the Netherlands have as their mother tongue. The answers were filled in the CAPI program in either Dutch or Moroccan Arabic. There was no need to translate questionnaires or advance letters for Surinamese or Antilleans. Dutch is the mother tongue for many, if not all persons of Surinamese or Antillean origin.

The second type of tailoring is the assignment of sample units to an interviewer with a shared ethnic background. In each design, all sampled persons of Moroccan or Turkish origin were contacted by a *bilingual* interviewer with a shared ethnic background during the face-to-face (and telephone) phase. In both the single- and mixed-mode design, about half of the sampled persons of Surinamese or Antillean origin in the telephone and/or face-to-face phase were approached by interviewers with a shared ethnic background. The other half of each sample was approached by either Dutch interviewers or interviewers with another ethnic background. The allocation of Surinamese and Antillean sample units to interviewers with a shared ethnic background was based on the availability of an interviewer with a shared ethnic background in the area.

3.2. Methods

A standard measure for judging the quality of a *response* sample is the response rate, despite the fact that it is not a direct measure and also a poor indicator of nonresponse bias (Biemer and Lyberg 2003; Groves and Peytcheva 2008). In the last few years, several other quality indicators have been developed that provide insight into the existence of nonresponse bias in survey estimates requiring somewhat weaker assumptions, such as *missing at random* (MAR) (Särndal 2011; Särndal and Lundström 2010; Schouten et al. 2009; Wagner 2010) or the weakest assumption, *missing not at random* (MNAR) (Andridge and Little 2011), and allow us to estimate its size. In order to answer our first research question – *overall quality* – we will use, next to the response rate, two approaches to evaluate how both designs affect the quality of the *response* samples and potential nonresponse bias in survey estimates for each design. In order to answer the second research question – *systematic differences* – differences in response propensity between sociodemographic subgroups, based on sample frame variables, are analyzed.

The First Approach for Assessing the Overall Quality (R1-1)

As a first approach for assessing the overall quality of the *response* samples, the representativity or R-indicator and the estimated maximal absolute *standardized* bias are used (Schouten et al. 2009). The R-indicator is a measure that describes how well the *response* sample reflects (i.e., how representative it is of) the population of interest, based on a certain number of background variables (Schouten and Cobben 2007; Schouten and Cobben 2008; Schouten et al. 2009). Obviously, this representativity only applies to the variables included in the model for estimating this measure and the response probability depends on these observed data only. One very important prerequisite is that the R-indicator needs complete (frame) data on all sample members: respondents and nonrespondents. This might not always be available. The R-indicator evaluates the differences in the estimated average response propensities between all strata, based on the variables included in the model from the available frame data. Response is considered representative if the response propensities are constant across the sample, which corresponds to a missing completely at random mechanism (Andridge and Little 2011, 154; Little and Rubin 2002).

Schouten et al. (2009, 107) show that “the R-indicator can also be used to set upper bounds to the non-response bias and to the root mean square error (RMSE) of adjusted response means.” The following equation (Eq. 1) from Bethlehem et al. (2011) shows the relation between the (estimated) average response probabilities ($\widehat{\overline{\rho}}$), the R-indicator $\widehat{R}(\widehat{\rho})$, the estimated standard deviation of the survey item $\widehat{S}(y)$, and the maximal absolute bias $\widehat{B_m}(\widehat{\rho}, y)$.

$$\widehat{B_m}(\widehat{\rho}, y) = \frac{(1 - \widehat{R}(\widehat{\rho}))\widehat{S}(y)}{2\widehat{\overline{\rho}}} \quad (1)$$

For an unambiguous comparison, Bethlehem et al. (2011) use the Cauchy-Schwarz inequality to factor out the S(y). This results in the estimated maximal absolute *standardized* bias (Eq. 2):

$$\widehat{B_m}(\widehat{\rho}, y) = \frac{\left(1 - \widehat{R}(\widehat{\rho})\right)}{2\widehat{\widehat{\rho}}} \tag{2}$$

The Second Approach for Assessing the Overall Quality (R1-2)

As a second approach for assessing the overall quality of the *response* samples the fraction of missing information estimates are used (Wagner 2008; 2010). The fraction of missing information (FMI) originates from the framework of multiple imputations (Dempster et al. 1977; Rubin 1987). It is a method used for incorporating uncertainty due to missing values in variance estimates and can be used to judge the efficiency of multiple imputations. FMI is defined as the ratio of the between-imputation variability to the total variance of the survey estimates (Wagner 2008; 2010).

The FMI is proposed as an alternative measure to the response rate to assess the quality of a sample with respect to potential nonresponse bias for a single item using all available data directly: complete case data plus paradata (sample frame data and process data) (Wagner 2008; 2010).

If the FMI is below the nonresponse rate it will serve as an alternative quality indicator to the response rate. Furthermore, provided we choose the correct model (i.e., the response probability depends only on the observed variables included in the model), it allows us to estimate the potential nonresponse bias for a specific survey item.

The $\widehat{B_m}\left(\widehat{\rho}, y\right)$ and the FMI approach differ in the way they estimate how nonresponse bias can impact the survey estimate. For instance, the $\widehat{B_m}\left(\widehat{\rho, y}\right)$ presented in Equations (1) and (2) is an estimate of the upper bound nonresponse bias for a hypothetical survey item, under the scenario where nonresponse correlates maximally to this variable (Schouten et al. 2011). It is based on the auxiliary variables in the model and an assumed correlation between these variables and the hypothetical survey item. There is no item-specific estimate for nonresponse bias.

Wagner's approach is designed to estimate the effect of nonresponse bias on the actual item level. In his approach, Wagner (2010) assumes that the missingness of the variable Y is independent of Y after conditioning on the covariates included in the model. This relates to a missing at random assumption (Andridge and Little 2011). Andridge and Little (2011) even extended the approach to MNAR models.

Given the difference in survey and item level-based estimates of nonresponse bias, it is interesting to compare the results of the $\widehat{B_m}\left(\widehat{\rho}, y\right)$ with the FMI approach to see whether they yield similar results. To this end we will compare the FMI results of multiple items and compare the combined results to the outcome of the $\widehat{B_m}\left(\widehat{\rho}, y\right)$.

Assessing Systematic Differences (R2)

Sometimes certain sociodemographic subgroups, such as young males, can be expected to have a different position or opinion on important research topics, such as having a job or the attitude on sociocultural integration. When they are under or overrepresented in the response sample, the results with respect to these research questions may be biased.

It is therefore important to see whether the different designs systematically affect the response composition of surveys among non-Western minorities and how they affect the response composition. To answer our second research question, to see whether the survey designs systematically cause different sociodemographic subgroups to be over- or underrepresented in the response samples among non-Western minority groups, partial R-indicators will be used (Schouten et al. 2011; Schouten et al. 2012; Shlomo et al. 2009).

These sociodemographic subgroups can be determined based on variables included in the model used to estimate the R-indicator. A partial R-indicator on a variable level shows the contribution of a specific background variable included in the model to the overall lack of representativity of the final sample. A partial R-indicator can also be calculated on a category level to ascertain the contribution to the lack of representative response separately for each category.

There are *unconditional* and *conditional* partial R-indicators for discrete variables and categories. The *unconditional* partial R-indicator on a variable level can be used to make comparisons between surveys (Shlomo et al. 2009, 7). It measures the variability of the response propensities between the different categories of a variable. The larger the variability, the greater the contribution to the lack of representativity. This indicator is non-negative and bounded above by 0.5 (Schouten et al. 2011, 236).

The values of the *unconditional* partial R-indicators on a category level may take values between −0.5 and 0.5 (Schouten et al. 2011, 236). A negative value indicates an underrepresented category and a positive value indicates an overrepresented category and zero (0) means representative.

The *conditional* partial R-indicator on a variable level measures the contribution of a variable to the lack of representative response, adjusted for the impact of the other variables included in the model (Schouten et al. 2011, 237). It tries to isolate the part of the nonrepresentative response that can be attributed to a specific variable. The conditional partial R-indicator on a variable level can take on any value in the interval [0, 0.5.]

The values of the *conditional* partial R-indicator on the category level range from 0 to 0.5 and show the conditional contribution of a category to the lack of representative response. The higher the value, the larger the contribution of the category to the lack of representativity.

4. Results of the Comparison of Single- and Mixed-Mode Designs Among Ethnic Minorities

4.1. Results on Overall Quality (R1-1): Representativity and the Maximal Absolute Standardized Bias

"When indicators are used to compare multiple surveys, and partial R-indicators could be part of such a comparison, then generally available auxiliary variables should be selected for which literature has shown that they relate to nonresponse in most if not all surveys" (Schouten et al. 2011, 15). In this section, the paradata used consists of the auxiliary sample frame variables *Age group*, *sex*, *municipality size* and *immigration generation*. All these variables have shown a large variability between the categories on the propensity to respond (see for instance Feskens et al. 2010; Groves and Couper 1998; Stoop 2005). No other complete frame data was available for inclusion in the analysis. The final R-indicator model we used consisted of *Age group* (six categories: 15–24; 25–34; 35–44; 45–54; 55–64; above 64 years); *Sex* (male and female); *Municipality size* (three categories: large, middle and small) and *Immigration generation* (first and second immigration generation), plus three interaction terms: *Age group* * *Municipality size*; *Immigration generation* * *Sex*; and *Immigration generation* * *Municipality size*.

For this study we used the AAPOR definition 1, the minimum response rate, to calculate the response rate (AAPOR 2011). Looking at the results in Table 2, the following pattern emerges. In each of the four mixed-mode samples a significantly higher response rate was achieved in comparison to their single-mode counterparts. However, the representativity of each of the single-mode *response* samples is significantly higher than each of the corresponding mixed-mode *response* samples. So, despite achieving the highest response rate, the mixed-mode *response* sample does not result in the best response composition with respect to the variables included in the model.

The $\widehat{B_m}$ takes into account both the response rate and the response composition with respect to the variables in the model (Eq. 2). The $\widehat{B_m}$ shows similar results to the R-indicator. The single-mode *response* samples all result in lower $\widehat{B_m}$ estimates than their mixed-mode counterparts.

The R-indicator shows that the SM design leads to a more representative sample compared to the MM design across and within ethnic groups, although there is no significant difference between the R-indicators of the Turkish SM and the Surinamese and Antillean MM design.

However, when the response rate is taken into account, resulting in the $\widehat{B_m}$ estimate, the SM design always leads to lower estimates for the upper bound nonresponse bias than the MM design-based estimates.

4.2. *Results on Overall Quality (R1-2): Fraction of Missing Information (FMI)*

The FMI was also used to assess how different survey designs affect the quality of the survey estimates. This was done separately for each of the four ethnic groups for both

Table 2. Response rate (RR_1), R-indicator ($\widehat{R}$), 95%-confidence interval R-indicator ($\widehat{R}^{Cl}_{0.95}$), maximal absolute standardized bias ($\widehat{B_m}$) and gross sample size (N′), separate for each ethnic group and survey design (single mode (SM) or sequential mixed mode (MM))

Ethnic group	Survey	RR_1 (%)	$\widehat{R}$ (%)	$\widehat{R}^{Cl}_{0.95}$ (%)	$\widehat{Bm}$(%)	N′
Turkish	SM	52.1	80.5*	(79.5–81.4)	18.8	1,564
	MM	54.5	76.8	(75.6–77.9)	21.4	9,78
Moroccans	SM	48.0	85.7*	(84.5–87.0)	14.8	1,737
	MM	51.7	75.8	(74.4–77.1)	23.4	1,086
Surinamese	SM	41.0	86.6*	(85.5–87.8)	16.4	1,929
	MM	43.1	80.7	(79.3–82.1)	22.4	1,203
Antilleans	SM	44.2	85.6*	(84.9–86.2)	16.4	1,973
	MM	44.4	79.1	(78.2–80.1)	23.4	1,231

Note: $*p = < 0.05$. N′ based on eligible cases.

designs. To estimate the FMI the following paradata were used: the same auxiliary variables (and interaction terms) from the sample frame as for the R-indicator plus the process data variable "number of contact attempts". Dummies were used to indicate contact via Web, CATI, one face-to-face contact attempt, two face-to-face contact attempts, and so on. Web was used as the reference category.

Since the FMI is an indicator of quality at the survey variable level and we want to evaluate the quality of both survey designs, we have selected and calculated the FMI for 16 different survey items. These items cover a wide range of topics (see Appendix A). The combined results should provide us with a good indication of the overall quality of the final response sample.

We followed the guidelines provided by Graham et al. (2007) and Wagner (2008) and we used 100 multiple imputations per item to reliably estimate the FMI separately for each ethnic group within each design. Table 3 presents the summary results of the analysis and the actual FMI estimates are shown in Appendix B.

In the SM design, the majority of the items included in the analysis have an FMI below the corresponding nonresponse rate (NR). This is true among all ethnic groups. This indicates that for the majority of the survey items included in the analysis, there is less uncertainty about the (mean) values for those estimates based on the imputed data compared to the estimates based on the complete case data only.

For the MM design the reverse is true, the FMI generally being above the corresponding nonresponse rate. This tells us that, using the same model, there is more uncertainty about the imputed values based on the MM survey data, which would indicate a less balanced sample. In this case the nonresponse rate is the better indicator for the survey data quality and the potential for nonresponse bias in a survey estimate than the difference between the response sample-based estimate and the estimate based on the fully imputed dataset.

There is a clear relationship between the (non)response rate and the fraction of missing information (see for instance, Wagner 2008). The higher the response rate, the lower the expected FMI. Within each ethnic group, the SM design resulted in a lower response rate

Table 3. Summary results of the fraction of missing information estimates ($\widehat{FMI}$) and for the 16 survey items, separately per ethnic group and survey design

	Turkish		Moroccans		Surinamese		Antilleans	
	SM	MM	SM	MM	SM	MM	SM	MM
No. of items with the $\widehat{FMI}$ below NR	14	4	12	4	14	0	13	0
No. of items with the lowest $\widehat{FMI}$ when SM and MM are compared within an ethnic group	14	2	12	4	16	0	16	0
No. of items in the SM for which the $\widehat{FMI}$ is below the MM NR rate compared within an ethnic group	12		12		14		12	

Note: FMI = fraction of missing information estimate; NR = nonresponse rate; SM = single-mode survey design; MM = sequential mixed-mode survey design.

than the MM design (see for instance Table 2). We could therefore have expected that within each group the FMI estimates based on the MM design would be below the FMI estimates based on the SM design. However, when compared within an ethnic group, the FMI estimates based on the SM survey data are mostly lower than the FMI estimates based on the MM survey data. Finally, the FMI estimates based on the SM design could still be above the nonresponse rate of the MM, because many of the MM FMI estimates were above their corresponding nonresponse rate. This means that the SM FMI estimates could still be surrounded by more uncertainty than the MM estimates based on the response rate. However, the majority of the FMI estimates based on the SM design are also below the nonresponse rate of the MM design within each ethnic group (Table 3, last row). All in all, these results can be seen as an indication that the single-mode design leads to better quality estimates across the ethnic groups than the sequential mixed-mode design. However, some caution is needed because the different modes in the sequential mixed-mode design may contribute additional uncertainty about the estimates based on imputed data due to mode-related effects (a model that included type of mode was also analyzed, but yielded similar results). Furthermore, we make the assumption that our model is correct and comparable within each separate ethnic group.

Comparison of the Estimated Maximal Absolute Standardized Bias ($\widehat{B_m}$) and the Mean of the 16 Fraction of Missing Information Estimates ($\widehat{\overline{FMI}}$)

Ideally both quality indicators should produce similar results because they incorporate response rate and the sample composition information and because more or less identical models were used to estimate both sets of indicators. To this end, we have compared the eight outcomes of $\widehat{B_m}$ with the eight outcomes of the $\widehat{\overline{FMI}}$ (plus standard deviation) to check whether or not they lead to similar conclusions (Table 4). We have chosen to use the $\widehat{\overline{FMI}}$ based on all 16 survey items to obtain an overall idea about the amount of uncertainty related to imputed means based on either SM or MM survey data.

The results differ somewhat if we compare both survey designs across all ethnic groups (Table 4). For instance, the lowest $\widehat{B_m}$ does not correspond with the lowest $\widehat{\overline{FMI}}$. Also, the four lowest $\widehat{B_m}$ estimates all come from SM *response* samples, whereas this is only true for three out of the four lowest values of the $\widehat{\overline{FMI}}$. However, the results are quite similar if we compare the indicators within an ethnic group. Within each ethnic group, both $\widehat{B_m}$ and $\widehat{\overline{FMI}}$ are lower when they are based on the SM data than on the MM data. This result makes sense because, while the $\widehat{B_m}$ is designed to be comparable across surveys, the predictive value of the auxiliary variables when used directly for imputation is most likely not the same for each sample. However, it will be much more similar in the two samples from the

Table 4. The estimated maximal absolute standardized bias ($\widehat{B_m}$), the mean and standard deviation of the 16 fraction of missing information estimates ($\widehat{\overline{FMI}}$) separately for SM and MM and ethnic group

	Turkish		Moroccans		Surinamese		Antilleans	
	SM	MM	SM	MM	SM	MM	SM	MM
$\widehat{\overline{FMI}}$(sd.)	44.7 (4.4)	51.0 (6.5)	50.1 (4.5)	53.3 (5.2)	54.0 (4.8)	70.2 (5.6)	49.7 (6.4)	61.4 (3.8)
$\widehat{B_m}$	18.8	21.4	14.8	23.4	16.4	22.4	16.4	23.4

same ethnic population. Still, we would gather that both estimates lead to the conclusion that the SM design outperforms the MM design.

4.3. Results on the Systematic Differences (R2): Partial R-Indicator Results

In order to answer our second research question, we want to find out whether there is a systematic impact of the survey design on the representativeness of the response across the auxiliary variable categories included in our response model. By 'systematic' we mean that the same pattern is seen across all ethnic groups. Accordingly we shall start by examining the evolution of the variation in response propensities for all variables included in the response model for the different stages of the sequential mixed-mode design, separately for each ethnic group. Next we will examine how the response samples at the different stages of the sequential mixed-mode survey compare to the response sample of the single-mode survey with respect to the variation of the response propensities.

In this section, the paradata used consists of the same four auxiliary sample frame variables. Table 5 shows the main findings of the (more or less) systematic impact that each separate mode in the sequential mixed mode had on the representativeness of the response for the variables included in our response model, separately for each ethnic group. The impact of CATI and CAPI in the sequential design shown here is conditional on the previous modes used. Also, the CATI and CAPI results refer to the unique impact and not the cumulative impact which is shown in Table 6.

Tables 5 and 6 also contain the main findings of the single-mode survey design, separately for each ethnic group. Appendix C contains the tables with the actual values of the unconditional and conditional partial R-indicators of these four variables. These tables contain the values of both the variable and category-level indicators of the various stages of the sequential mixed-mode *response* samples and the single-mode CAPI *response* samples, separately for each ethnic group.

For ease of interpretation the different stages of the sequential mixed-mode design are presented first, followed by the single-mode design (SM), separately for each group. Rows indicated with "++++" mean a consistent pattern of overrepresentation across ethnic groups of the sociodemographic category within a certain survey mode. Rows indicated with "− − − −" mean a consistent pattern of underrepresentation across ethnic groups of the sociodemographic category within a certain survey mode. Rows indicated with a combination of "+" and "0" (e.g., ++ 0 0) mean a mostly consistent pattern of representative to over representative response across ethnic groups of the socio-demographic category within a certain survey mode. Rows indicated with a combination of "−" and "0" (e.g., − − 0 0) mean a mostly consistent pattern of underrepresentative to representative response across ethnic groups of the sociodemographic category within a certain survey mode. Finally, empty rows indicate that no consistent pattern can be discerned across ethnic groups of the sociodemographic category within a certain survey mode.

The Introduction of WEB (M_{web})

The use of WEB causes differing levels of representativeness with respect to the variables included in the response model across the four ethnic groups. *Age group* and *immigration*

Table 5. Systematic impact of each separate stage in the sequential mixed-mode design and the single-mode design on the representative response of the variables included in the response model, separately for each ethnic group

	M_{web}				M_{tel}				M_{f2f}				SM			
	T	M	S	A	T	M	S	A	T	M	S	A	T	M	S	A
Age group																
15-24	+	+	+	+					+	+	+	+	+	+	+	+
25-34					–	–	–	–	–	–	–	–	–	–	–	–
35-44																
45-54	–	–	0	–									+	+	+	+
55-64	–	–	–	0	0	+	+	+					+	0	0	0
>64	–	–	0	–	+	0	+	+	+	+	+	+	+	0	+	0
Gender																
Male																
Female					+	+	+	+								
Municipality size																
Large	0	–	–	–									–	0	–	–
Midsize													+	+	+	+
Small	0	0	0	+	+	+	+	0								
Immigration generation																
1st generation	–	–	–	–	0	0	+	+	+	0	+	+				
2nd generation	+	+	+	+												

Note: M_{web} = result of the introduction of WEB; M_{tel} = result of the introduction of CATI in the mixed-mode sequence; M_{f2f} = result of the introduction of CAPI in the mixed-mode sequence; S = result of the single mode; T = Turkish; M = Moroccans; S = Surinamese; A = Antilleans; '+' = overrepresented; '-' = underrepresented '0' = representative. +, 0 and – are based on whether or not zero is included in the approximated confidence interval.

Table 6. Overview of the systematic impact the different stages of the sequential mixed-mode design have on the variation in the response propensities of the variables included in the model compared to the single-mode design, separately for each ethnic group

	MM WEB vs. SM								MM WEB + CATI vs. SM								MM vs. SM							
	MM WEB				SM				MM WEB + CATI				SM				MM				SM			
	T	M	S	A	T	M	S	A	T	M	S	A	T	M	S	A	T	M	S	A	T	M	S	A
Age group																								
15-24	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
25-34					–	–	–	–					–	–	–	–	–	–	–	–	–	–	–	–
35-44																								
45-54	–	–	0	–	+	+	+	+	–	–	0	–	+	+	+	+					+	+	+	+
55-64	–	–	–	0	+	0	0	0					+	0	0	0	0	–	–	0	+	0	0	0
>64	–	–	0	–	+	0	+	0					+	0	+	0					+	0	+	0
Gender																								
Male									–	–	–	–					–	–	–	–				
Female									+	+	+	+					+	+	+	+				
Municipality size																								
Large	0	–	–	–	–	0	–	–	–	–	–	–	–	0	–	–	–	–	–	–	–	0	–	–
Midsize					+	+	+	+	+	+	+	0	+	+	+	+	+	+	+	+	+	+	+	+
Small	0	0	0	+					+	+	+	+												
Immigration generation																								
1st generation	–	–	–	–					–	–	–	–					–	–	–	–				
2nd generation	+	+	+	+					+	+	+	+					+	+	+	+				

Note: T = Turkish; M = Moroccans; S = Surinamese; A = Antilleans; '+' = overrepresented; '-' = underrepresented; '0' = representative. +, 0 and – are based on whether or not zero is included in the approximated confidence interval.

generation show a strong collinear response behavior among the Turkish and the Moroccans (see unconditional and conditional partial R-indicators in Appendix C). This was to be expected, since Turkish and Moroccan immigration only started in the mid-1960s and therefore second-generation immigrants over the age of 45 hardly exist (CBS-Statline). The first-generation immigrants were mostly men who came to the Netherlands for work. Partner reunification only started in the mid-seventies. Our data suggest that across all ethnic groups the young (15–24) and second-generation sampled persons find it easier to respond via WEB. The older (45 upwards) and first-generation sampled persons seem to be systematically underrepresented. Furthermore, there is also a systematic effect of WEB across the ethnic groups when it comes to *municipality size*. Persons from large cities are less inclined to participate via WEB. Finally, the use of WEB does not appear to have a systematic impact on *gender* across the ethnic groups.

The Introduction of CATI in the Sequence (M_{tel})

The success of the CATI mode was quite limited, resulting only in a very modest increase in response across the ethnic groups. Therefore the introduction of CATI in this sequence had a limited impact on the representativeness of response for the variables included in the response model. However, CATI does attract a very selective response group. The use of CATI in this sequence mainly results in female respondents, older respondents, first-generation respondents and respondents who live in small municipalities.

The Introduction of CAPI in the Sequence (M_{f2f})

The introduction of CAPI as the final mode of contact in the sequential mixed-mode design has a systematic effect on *age group* and *immigration generation* across the ethnic groups compared to WEB+CATI. With respect to *age group,* the face-to-face interviewers get either young (15 to 24) and/or older (above 64) persons to respond, but fail to get persons in the age of 25 to 34 to respond. Finally, face-to-face interviewers are able to get first generation immigrants to respond across all ethnic groups. Interestingly enough, there seems to be no systematic effect for *gender* or *municipality size* when CAPI is introduced as the final mode in this sequence.

SM: the Use of CAPI Only

The use of CAPI as a single mode of surveying ethnic minorities has a strong impact on the way different age categories are represented in the response. Persons aged 25 to 34 do not respond well and are underrepresented across all ethnic groups. The SM design also systematically results in an overrepresentation of persons aged 15 to 24. With respect to the upper three age categories, the SM design also causes these categories to be somewhat overrepresented, rather than representative response or an underrepresentation across all ethnic groups.

The SM design results in a systematic overrepresentation of persons living in midsize cities. It also leads to an underrepresentation of persons living in large cities, although among Moroccans the response is more or less representative. Finally, the SM design did not seem to have a systematic effect on *gender* or *immigration generation* across the different ethnic groups.

Partial R-Indicator Comparison Between the Different Survey Designs

The partial R-indicators on the variable level show some significant differences in the variation of the response propensities for the variables included in the response model (see Appendix C). This means that the use of different survey designs (or intermediate mode combinations of the MM design) causes different response compositions and that the size of the variation in response propensities is dependent on ethnic group, mode and variable. For instance, the use of WEB does not lead to a larger variation of the response propensities than the SM design for all the variables included in the response model, but it is dependent on the interaction between the response variable and ethnic group.

The differences in the variation of response propensities between different survey designs can also be the result of the same sociodemographic categories being more heavily under or overrepresented. For example, both the WEB and SM samples result in an overrepresentation of persons aged 15 to 24, but they differ in the degree of overrepresentation.

In order to gain a better understanding of the advantages and disadvantages of (combinations of) the current sequential mixed-mode survey design compared to a single-mode CAPI survey design, the results of the former are compared to the results of the latter in a more detailed manner.

For this comparison we will focus on whether the different survey designs cause the same or different sociodemographic categories to be systematically over- or under-represented across ethnic groups or whether this is dependent on ethnic group.

MM WEB Versus SM

The first step of the MM design (WEB only) and SM design causes some of the same categories to be under- or overrepresented (Table 6). For instance, both result in an overrepresentation of persons aged 15 to 24. Secondly, both mostly result in a small to rather large underrepresentation of big city dwellers and a representative response or overrepresentation of persons from midsize municipalities.

WEB only and the SM design also lead to the systematic under- or overrepresentation of different categories across all ethnic groups. The use of WEB usually results in an underrepresentation of the upper age categories, whereas the use of the SM design more often results in an overrepresentation of the upper age categories. Furthermore, the SM design systematically leads to an underrepresentation of persons aged 25 to 34, whereas for WEB this depends on the ethnic group. Furthermore, the use of WEB leads to a systematic underrepresentation of first-generation immigrants, which is not the case in the SM design.

An interesting result is the absence of a systematic impact of WEB only and the SM design for *gender* across the ethnic groups. As it turns out, both WEB only and the SM design lead to an over- or an underrepresentation of males (or females), dependent on ethnic group.

MM WEB+CATI Versus SM

The use of CATI as a second step in the mixed-mode sequence resulted in a low response and is therefore not recommended for ethnic minority groups. As a result of the low response rate, the impact on the response composition is rather small and marked by the same differences and similarities found in the WEB versus SM comparison. However, because of the very selective response group in CATI, the systematic differences between WEB+CATI and the SM design have decreased somewhat for the upper age categories. Furthermore, the WEB+CATI design leads to a systematic underrepresentation of men and systematic overrepresentation of women, as opposed to the SM design.

MM Versus SM

The samples of the complete MM design show some interesting similarities with the SM design across the ethnic minorities. Both designs lead to a systematic overrepresentation of persons aged 15 to 24 and an underrepresentation of persons aged 25 to 34. They also yield the same sort of result when it comes to *Municipality size*. They both result in a systematic underrepresentation of big city dwellers and an overrepresentation of persons from midsize municipalities.

Both designs also lead to some systematic differences with respect to sociodemographic categories. First of all, the upper age categories systematically tend to be somewhat overrepresented in SM, whereas this is not a systematic finding in the MM. The opposite is actually true for persons aged 55 to 64. There is a tendency for this age group to be underrepresented in the MM. The MM design also results in an underrepresentation of men and first-generation immigrants, as opposed to the SM design. However, the underrepresentation of first-generation immigrants in MM is less severe than in the WEB+CATI design.

4.4. The Cost Perspective

The use of a sequential mixed-mode design instead of a single-mode CAPI design has the potential to greatly reduce the costs of the survey. Theoretically, the largest cost savings are made when the sequential mixed-mode design introduces the most inexpensive mode (web or postal) first and follows up with increasingly more expensive, interviewer-assisted modes. Furthermore, this can generate economies of scale when the sample size increases.

However, there are costs and cost-related considerations which are either unique or amplified in case of a sequential mixed-mode design as compared to a single-mode CAPI design that easily can be overlooked. These are especially relevant when sample sizes are relatively small and the known survey difficulties in connection with specific target populations require the use of a CAPI mode.

First of all, there are the extra costs related to questionnaire development and interviewer training. These costs can increase because the questionnaire has to be developed to be suitable for every mode and administered in different interviewer-assisted modes. From this point of view, CATI is not very cost effective as a mode among non-Western minorities in this design: only 1.3% to 6% of the sampled persons in the different ethnic groups responded via CATI.

Secondly, information costs money and, compared to a face-to-face survey design, the use of a sequential mixed-mode design limits the amount of information that can be gathered. In this experiment, the WEB and CATI questionnaire was reduced to about two-thirds of the length of the CAPI questionnaire. This means that the cost per survey question can actually increase in a sequential mixed-mode survey.

Thirdly, time is money: the length of the fieldwork period can increase because of the use of a sequential mixed-mode design. Each mode needs a certain amount of time to be used to its full potential. For instance, in this study the second mode (CATI) was only introduced one and a half months into the fieldwork period. The need to wait for each mode to reach its full potential was the main reason for which the reissue in the sequential mixed-mode design had to be cut short. In addition, there are logistic costs related to conducting a sequential mixed-mode survey. It needs to be monitored quite carefully if and when a nonresponding sampled person can 'move' from one mode to the next.

Fourthly, there is a potential for a relative increase in travel costs for face-to-face interviewers. From a logistic point of view, the remaining number of nonresponding sampled persons in the CAPI phase of the MM design can be inconveniently located. This can also cause a reduction in the number of contact attempts an interviewer is able to conduct in a single day. It goes without saying when an interviewer is working on several surveys at the same time, this might not pose a problem.

A fifth, mixed-mode related cost concerns interviewer motivation and effort per face-to-face interview. Table 7 shows the ratio between the number of interviews and the total number of contact attempts conducted in the CAPI mode, separately for each ethnic group and survey design.

The ratio of face-to-face contact attempts to number of interviews is substantially higher in the MM compared to the SM. For instance, among the Turkish, for each 4.5 contact attempts that were made in the SM design, there was one interview completed, whereas in the MM design, this ratio was 5.3 to 1. Furthermore, the ratio among the Turkish and the Moroccans is a lot lower than among the Surinamese and the Antilleans. This indicates that a lot more unsuccessful contact attempts took place among the Surinamese and the Antilleans. This results not only in a lower response rate, but also in more effort per interview.

Put simply, face-to-face interviews are more expensive in terms of return when they are conducted as part of a sequential design. This result is of course to be expected since the 'easy' respondents have already participated via WEB or CATI, leaving the more reluctant or hard to reach sampled persons. However, the estimated costs of a face-to-face interview are to some extent based on the number of unsuccessful contact attempts that are made for each successful contact attempt. Therefore, the increased amount of effort needed in the MM CAPI phase when comparing the costs of a CAPI interview in a single-mode survey to a CAPI interview in a mixed-mode survey should be taken into account. This result not only has a direct financial implication; it can also lead to decreased motivation among interviewers, which in turn might lead to additional costs (bonus arrangements) or an

Table 7. Ratio of face-to-face contact attempts to number of interviews conducted in the CAPI mode during the first fieldwork phase for the SM and the MM samples, separately for each ethnic group

	Turkish		Moroccans		Surinamese		Antilleans	
	SM	MM	SM	MM	SM	MM	SM	MM
Ratio	4.5	5.3	3.9	5.8	10.6	13.8	10.1	12.4

extension of the fieldwork period due to interviewers dropping out due to lack of motivation.

A final cost concern is related to analysis. It should not be forgotten that a sequential mixed-mode design will cost additional analysis time in order to check and correct for potential mode effects that can distort the results.

The eventual cost savings in this experiment, generated by using the current sequential mixed mode design instead of a single-mode face-to-face design among ethnic minority groups, amounted to between 12 to 20%, depending on how one would distribute fixed costs between both designs. However, given that this design choice also resulted in less information on the population of interest, a longer fieldwork period, additional analysis time and greater uncertainty related to the survey estimates based on both quality indicators, it can be concluded that in this instance the cost savings did not outweigh the reduction in quality.

5. Conclusion and Discussion

In this article we investigated how the use of a sequential mixed-mode – WEB-CATI-CAPI –design affects the quality of the *response* sample compared to a single-mode face-to-face CAPI design in surveys among non-Western minority groups in the Netherlands, as well as how these different survey designs may impact nonresponse bias on survey estimates. Statistics Netherlands drew two random samples from each of the four largest non-Western minority populations living in the Netherlands. In each ethnic group, one sample was assigned to a sequential mixed-mode design and a one sample to single-mode face-to-face CAPI design. This resulted in eight samples for analysis.

Furthermore, we analyzed whether the different survey designs enhance response rates to different degrees among different sociodemographic subgroups based on auxiliary variables. We also discussed costs and cost-related issues particular to this sequential mixed-mode design that are relevant in the quality versus costs trade-off decision.

Besides the response rate, we used two approaches to evaluate the quality of the *response* samples and potential nonresponse bias in survey estimates for both survey designs among non-Western minorities. The first approach was the representativity indicator (R-indicator) and the maximal absolute standardized bias ($\widehat{B_m}$) proposed by Schouten et al. (2009). The second approach was the fraction of missing information (FMI) proposed by Wagner (2008).

The sequential mixed-mode design resulted in higher response rates than the single-mode CAPI design in each of the four non-Western minority groups. However, both the R-indicator and the FMI approach showed that the single-mode CAPI survey design resulted in better quality *response* samples among non-Western minorities than the sequential mixed-mode survey design. Furthermore, the result of both the $\widehat{B_m}$ and the mean FMI analyses indicated that the potential for nonresponse bias in survey estimates is higher among the final samples based on a sequential mixed-mode design.

An analysis of partial R-indicators on the variable and category level was carried out to find out whether the survey designs enhance response rates differently among different sociodemographic subgroups. Overall, the variations in response propensities are larger in the sequential mixed-mode design than in the single-mode design for the variables included in the model, with *age group* and *municipality size* showing the largest contributions.

The partial R-indicator analysis also showed that the sequential mixed-mode design systematically resulted in an underrepresentation of men, persons aged 55 to 64 and first-generation immigrants across all ethnic groups, but this pattern was not repeated for the single-mode survey design. On the other hand, the single-mode CAPI survey resulted in an overrepresentation of persons from the upper age categories (45+) among all ethnic groups, which was not the case for the sequential mixed-mode design. Furthermore, both survey designs systematically caused an underrepresentation of persons aged 25 to 34 as well as big city dwellers and an overrepresentation of young persons (15 to 24) and respondents from middle size municipalities. This systematic impact of the different survey designs on the response composition is important to bear in mind when a strong correlation is expected between a survey topic and specific over- or underrepresented sociodemographic subgroups.

The impact of each mode in the sequential mixed-mode design on the response composition was also assessed. WEB is a good startup mode to survey ethnic minorities, but cannot be recommendable as the only mode. WEB mostly results in response from young persons and second-generation immigrants across all ethnic groups.

CATI is not very suitable as a follow-up mode for conducting a survey among ethnic minorities in the Netherlands and should be avoided. It leads to a selective and low response due to high rates of refusals and non-contact. Furthermore, penetration rates are very low across the ethnic groups, especially if CATI is used as a second mode. Only 10 to 25% of the WEB nonresponders could be matched to a known phone number (Korte and Dagevos 2011).

CAPI remains a necessary part of any survey of non-Western minorities in the Netherlands. The introduction of CAPI in the sequential mixed-mode design increases the response among young and old (> 64) persons and first generation immigrants across all ethnic groups.

The cost savings of 12 to 20% with the current mixed-mode design did not justify the decrease in *response* sample quality as indicated by the R-indicator, $\widehat{B_m}$ and FMI. This design choice not only resulted in a lower-quality *response* sample and greater uncertainty related to the survey estimates in terms of nonresponse bias, but it also resulted in additional ‘costs’ in terms of loss of information due to shorter questionnaires, extended fieldwork time, and extra analysis time. These and other cost-related issues, such as the costs in terms of development, effort, and support versus return for the different modes and additional monitoring should be carefully reviewed before the decision to make use of a sequential mixed-mode design. Especially for relatively small sample sizes and known survey difficulties in connection with specific target populations, these additional costs may outweigh the expected savings.

The mixed-mode results do provide insight into how to improve the quality of the sample for surveys among ethnic minorities, while possibly reducing costs. A sequential WEB+CAPI design with a complete reissue or even targeted re-issue of nonresponding sample units from underrepresented sociodemographic subgroups seems better suited to yield a high and balanced response among ethnic groups than the current sequential mixed mode design, while also reducing the length of the fieldwork period. This is the case provided the need for information does not exceed the optimal length of a WEB questionnaire. Furthermore, this design would still be less expensive to execute than a single mode CAPI design with a complete or targeted re-issue. In the re-issue, the nonresponding sampled persons should be assigned to other interviewers. To reduce the costs even more, one could consider reducing the number of face-to-face contact attempts to three or four during the first phase of fieldwork (Kappelhof 2014).

There are also several limitations to the current study. First of all, there are assumptions that go with the quality indicators used to assess the potential for nonresponse bias on survey estimates. Both quality indicators make use of the MAR assumption which is quite a strong assumption. Furthermore, in case of the R-indicator and the related measure of maximal absolute bias, no direct nonresponse bias estimate is possible since these measures are developed to compare surveys. In the case of the quality indicator based on the FMI approach, it is possible to provide direct estimates of nonresponse bias for a survey estimate given the MAR assumption. However, these results were not provided since the possibility of increased measurement variability because of the use of different survey modes in the sequential mixed-mode survey would distort the results too much (i.e., how much of the observed difference between the estimate based on the response rate and the imputed estimate was the result of nonresponse bias and how much can be contributed to the increased measurement variability). As a result, only the FMI estimates were presented as indicators of possible nonresponse bias occurrence in survey estimates. However, even then we have to assure ourselves that the measurement errors are the same across all response rates. If not, then comparing patterns of nonresponse across two designs without looking at the measurement errors is not as useful.

Another argument against our approach for estimating the FMI is that it is not actually necessary to fit the same model (i.e., include the same variables) to obtain the FMI of each dependent variable in order to be able to compare both designs. One may need a different set of predictor variables to obtain the best prediction for each separate dependent variable. Furthermore, as Andridge and Little (2011) argue, predictors used to predict response may differ from the predictors used to predict the outcome of substantive variables. Thus, it may be worth also considering other models to estimate and compare the FMI estimates which may lead to different results. However, our results are very consistent across ethnic groups and across different variables and present a fairly convincing picture that the response to MM design is highly selective for these specific populations. Nevertheless, future research should include several competing, but plausible (i.e., include variables known to correlate with the outcome variable) models to investigate to what extent the results are robust.

Finally, an interesting extension on the current study would be to include a quality indicator that allows for a direct estimate of nonresponse bias, but for which the model used for the estimates is based on the least restrictive assumption (MNAR), such as the proxy pattern-mixture approach of Andridge and Little (2011). This would allow for even more direct information that can be used in the cost- versus quality trade-off decision concerning which survey design is best suited to survey minority ethnic populations given financial and time restrictions.

Appendixes

Appendix A. Overview of the 16 Survey Questions Used in the FMI Approach

1	Do you see yourself as <ethnic group>? (Yes: No)
2	Are you currently employed? (Yes: No)
3	Do you consider yourself to be a member of a certain religion? (Yes: No)
4	To what degree do you consider yourself to be happy? (5-point scale)
5	Do you feel more <ethnic group> or Dutch? (5-point scale)
6	Generally speaking, how would you rate your health? (5-point scale)
7	Do you or your parents rent or own the house you live in? (rent/own/other)
8	Have you been discriminated against by native Dutch? (5-point scale)
9	In the Netherlands you get offered all the opportunities (5-point scale)
10	Do you have children? (Yes/No)
11	How satisfied are you with the Dutch society? (10-point scale)
12	How often did you visit a MD for yourself in the last two months? (0 to 60)
13	Do you own or have access to a computer to use for internet? (Yes/No)
14	It is better if the man is responsible for the finances (5-point scale)
15	How often do you experience difficulties when you have to talk in Dutch? (do not speak Dutch, often, sometimes or never)
16	How often did you do sports in the last 12 months?

Appendix B. Fraction of Missing Information Estimates (FMI in %) and the Nonresponse Rate (NR in %) for the 16 Survey Items, Separately for Each Ethnic Group and Survey Design (SM and MM)

	Turkish		Moroccans		Surinamese		Antilleans	
	SM	MM	SM	MM	SM	MM	SM	MM
$FMI_{Ethnic\ self}$	44.5	51.4	46.0	51.6	51.2	69.9	44.2	60.1
$FMI_{Employment}$	43.9	41.2	48.2	48.3	49.8	66.0	42.9	56.9
$FMI_{Religious}$	43.8	52.4	48.2	45.6	54.4	68.5	41.0	60.2
$FMI_{Happiness}$	50.7	56.1	51.3	58.1	56.7	75.0	47.3	65.6
$FMI_{Self\text{-}identification}$	53.9	63.0	58.0	56.0	56.9	74.2	56.4	57.0
FMI_{Health}	41.7	53.2	48.4	55.5	55.8	74.1	47.8	65.3
FMI_{House}	45.8	49.6	53.4	50.5	53.0	65.4	47.9	57.5
$FMI_{Discrimination\ self}$	45.3	51.2	51.7	55.9	50.9	74.7	61.5	63.5
$FMI_{Opportunities}$	47.1	56.8	55.7	57.7	56.2	72.6	60.3	61.6
$FMI_{Children}$	36.7	40.9	44.3	43.3	45.1	59.6	42.9	57.8
$FMI_{Satisfaction\_Society}$	47.5	59.5	54.5	57.6	61.4	77.1	57.4	70.8
FMI_{MD}	44.6	52.0	52.6	58.3	64.2	70.2	47.8	62.4
$FMI_{Internet}$	42.7	44.1	48.1	52.3	48.0	70.6	50.5	57.5
$FMI_{Man\_finance}$	45.1	52.6	52.4	59.8	55.7	77.0	53.5	62.6
FMI_{Speak_Ductch}	36.2	51.1	40.1	56.1	51.4	58.5	48.4	61.8
$FMI_{Sports\_frequency}$	45.3	40.6	48.1	46.4	53.8	69.1	45.3	61.1
NR	48.0	45.5	52.0	48.3	59.1	56.9	55.8	55.6
$NR_{Self\_identication^{1}}$	48.0	46.0	52.5	49.3	59.8	57.9	56.7	56.2
$NR_{House^{1}}$	48.4	46.3	53.3	48.8	59.1	56.9	56.0	56.1
$NR_{Discrimination\_self^{1}}$	48.0	46.2	53.1	49.0	59.1	57.1	56.2	56.3
$NR_{Opportunities^{1}}$	48.1	46.3	52.7	49.3	59.6	57.9	56.6	56.9
$NR_{Satisfied\_Society^{1}}$	48.2	45.7	52.6	48.6	59.2	57.6	55.9	55.8
$NR_{MD^{1}}$	49.2	47.2	54.1	51.6	59.3	58.6	55.9	57.0
$NR_{Man\_finance^{1}}$	48.1	45.6	52.6	49.3	59.1	57.4	56.1	55.9
N	1564	978	1737	1086	1929	1203	1973	1231

Note: [1] is corrected for item nonresponse.

Appendix C

Table C1. The unconditional variable- and category level partial R-indicators (multiplied by 1000) and response rate (RR_1), representativity indicator ($\widehat{R}$) and maximal absolute standardized bias ($\widehat{B_m}$) in %, separate for each ethnic group and survey design stage

	Turkish				Moroccans				Surinamese				Antilleans			
	MM WEB	MM WEB +Cati	MM	SM	MM WEB	MM WEB +Cati	MM	SM	MM WEB	MM WEB +Cati	MM	SM	MM WEB	MM WEB +Cati	MM	SM
Unconditional																
Age group	60.8	53.1*	79.0	51.4[ac]	60.7	60.4*	87.7	20.0[abc]	34.0	27.9*	48.9	29.8[c]	37.8	35.4	34.8	34.8
15-24	43.9	39.5	61.6	21.1	35.4	37.9	70.0	8.5	23.0	12.0	18.2	11.4	18.5	8.5	18.5	9.3
25-34	7.0	−5.9	−44.0	−36.0	17.2	12.7	−11.0	−16.6	−12.2	−20.4	−34.4	−16.1	−6.2	−12.3	−26.8	−28.7
35-44	−9.2	2.9	−1.3	−13.1	−9.5	−13.5	−21.6	4.7	3.8	5.6	−0.8	−12.2	9.2	11.6	11.4	9.5
45-54	−23.1	−18.6	−22.3	13.8	−20.6	−17.6	−7.3	5.4	0.8	1.1	5.0	4.8	−21.3	−18.5	−3.4	14.3
55-64	−16.5	−19.5	2.8	14.9	−21.1	−12.6	−29.7	−0.5	−21.3	−8.4	−8.9	2.1	4.8	19.0	−2.0	−2.2
64+	−28.8	−22.1	−1.0	17.6	−34.2	−37.4	−35.6	−0.0	−1.7	10.6	27.6	17.9	−21.9	−13.6	−2.4	1.8
Sex	4.1*	16.0	18.3	37.6[abc]	37.3	37.5*	17.9	6.9[abc]	2.2*	18.4	27.9	4.6[bc]	27.9	33.8*	17.4	11.0[ab]
Male	2.8	−11.0	−12.7	−26.2	−26.1	−26.3	−12.5	4.8	−1.6	−13.3	−20.2	3.2	−19.9	−24.0	−12.4	−7.6
Female	−3.0	11.5	13.2	26.9	26.6	26.8	12.8	−5.0	1.5	12.7	19.2	−3.1	19.6	23.7	12.2	8.0
Municipality	4.0*	16.9	17.0	26.3[a]	18.9	17.6*	51.2	15.8[c]	40.6*	49.1*	31.5	44.9[c]	39.1	37.8*	80.7	45.6[c]
Large	−2.2	−13.1	−13.7	−4.7	−13.7	−11.7	−31.6	0.1	−25.6	−32.7	−16.5	−24.5	−21.3	−23.3	−65.0	−33.1
Medium	−0.1	10.6	8.7	16.5	11.8	5.3	10.3	8.2	31.6	36.1	25.3	7.9	−1.3	2.2	43.9	31.0
Small	3.3	13.3	5.0	−20.0	5.1	12.1	38.9	−13.5	−0.7	6.1	−8.7	36.8	32.7	29.6	19.0	−4.4
Immigration generation	52.0	44.6*	33.0	32.0[ab]	51.6	54.2*	65.4	17.5[abc]	41.8*	25.0*	14.1	1.1[abc]	63.7*	48.1*	18.2	3.4[abc]
1G	−29.7	−25.4	−18.9	18.3	−29.8	−31.3	−37.8	−9.7	−25.3	−15.1	−8.5	−0.6	−33.6	−25.3	−9.6	−1.8
2G	42.7	36.6	27.1	−26.2	42.1	44.2	53.3	14.6	33.3	19.9	11.2	0.9	54.1	40.9	15.4	2.9
RR_1	21.4	25.3	54.5	52.1	22.7	24.0	51.7	48.0	20.9	26.9	43.1	41.0	21.6	26.2	44.4	44.2
$\widehat{R}$	84.6	84.9	76.8	80.5	83.0	82.0	75.8	85.7	83.9	82.1	80.7	86.6	80.4	81.4	79.1	85.6
$\widehat{B_m}$	36.0	29.8	21.4	18.8	37.4	37.5	23.4	14.8	38.5	33.3	22.4	16.4	45.4	35.5	23.4	16.4
N[1]	978	978	978	1564	1086	1086	1086	1737	1203	1203	1203	1929	1231	1231	1231	1973

Note. A significant difference ($p = < 0.05$) between mode and subsequent mode within SIM MM is noted with an *; [a] = significant difference between MM WEB only and SIM SM within ethnic group; [b] = significant difference between MM WEB+CATI and SIM SM within ethnic group; [c] = significant difference between SIM MM and SIM SM within ethnic group. [d] = significant difference between different ethnic groups. Standard errors were approximated (not included here) using 1000 bootstrap replicates of the estimates. [1] Based on all eligible cases.

Table C2. *The conditional variable- and category level partial R-indicators (multiplied by 1000), separate for each ethnic group and survey design*

	Turkish				Moroccans				Surinamese				Antilleans			
	MM WEB	MM WEB +Cati	MM	SM	MM WEB	MM WEB +Cati	MM	SM	MM WEB	MM WEB +Cati	MM	SM	MM WEB	MM WEB +Cati	MM	SM
Conditional																
Age group	35.6	33.9	71.8	60.5	36.2	34.0	60.6	23.8	32.7	35.8	52.2	31.1	27.8	34.0	29.6	37.6
15-24	19.4	17.8	46.7	39.7	11.3	10.6	39.7	6.2	6.5	7.2	14.1	11.6	2.3	4.9	15.4	10.5
25-34	15.5	15.6	50.2	36.3	17.8	13.1	31.4	18.3	22.1	27.5	38.4	15.6	5.4	10.2	22.4	30.6
35-44	10.4	14.5	13.0	22.9	8.3	6.3	12.5	9.8	12.6	12.6	6.0	13.9	13.3	15.5	10.7	10.3
45-54	10.2	7.5	14.0	6.5	4.5	2.1	11.8	9.2	14.0	11.0	10.5	4.9	11.3	11.8	1.9	15.6
55-64	6.6	12.0	7.7	7.8	13.6	4.1	17.0	2.9	12.8	5.2	6.9	3.0	12.2	23.0	3.3	3.2
64 +	20.0	13.5	5.9	11.8	24.3	26.5	22.8	2.1	4.0	12.9	29.2	19.0	16.9	11.1	3.6	2.4
Sex	2.6	17.5	20.2	36.5	33.9	34.9	15.6	5.3	4.0	19.7	29.0	5.2	31.5	34.9	17.6	11.9
Male	1.8	12.0	14.0	25.4	24.0	23.7	11.0	3.7	2.9	14.2	21.0	3.7	22.1	24.7	12.5	8.1
Female	1.9	12.6	14.5	26.1	23.9	23.6	11.2	3.8	2.8	13.6	20.0	3.6	22.4	24.7	12.4	8.6
Municipality	5.0	11.8	12.9	30.5	22.9	18.8	51.5	15.3	38.0	46.6	30.5	45.5	33.7	31.4	77.8	46.8
Large	1.6	9.5	10.2	8.3	15.8	13.5	34.2	1.1	23.7	30.8	15.4	25.3	17.6	19.1	62.2	32.8
Medium	3.3	6.4	7.1	20.3	16.4	9.1	15.6	7.9	29.5	34.4	24.4	9.1	5.0	4.0	43.2	32.5
Small	3.3	2.8	3.1	21.1	2.2	9.3	35.2	13.0	3.3	6.0	9.7	36.7	28.3	24.6	17.8	7.7
Immigration generation	16.5	16.0	2.0	45.0	17.0	21.0	9.2	21.6	37.2	29.9	20.6	1.4	54.1	45.1	7.3	3.1
1G	11.6	11.0	1.3	32.2	12.2	15.2	6.5	15.4	24.5	19.7	14.1	1.0	33.2	27.5	4.4	1.9
2G	11.7	11.6	1.5	31.4	11.8	14.5	6.5	15.2	27.9	22.4	15.0	1.0	42.7	35.8	5.8	2.4

6. References

AAPOR 2011. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. (7th edition). The American Association for Public Opinion Research. Available at: http://www.aapor.org/AM/Template.cfm?Section=Standard_Definitions2&Template=/CM/ContentDisplay.cfm&ContentID=3156 (accessed December 2013).

Andridge, R.R. and R.J. Little. 2011. "Proxy Pattern-Mixture Analysis for Survey Nonresponse." *Journal of Official Statistics* 27: 153–180.

Bethlehem, J., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Biemer, P.P. and L.E. Lyberg. 2003. *Introduction to Survey Quality*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Bijl, R. and A. Verweij. 2012. *Measuring and Monitoring Immigrant Integration in Europe: Integration Policies and Monitoring Efforts in 17 European Countries*, 2012-8. Den Haag: SCP. Available at: http://www.scp.nl/english/Publications/Publications_by_year/Publications_2012/Measuring_and_monitoring_immigrant_integration_in_Europe (accessed June 2013).

CBS-Statline. Available at: http://statline.cbs.nl/StatWeb/publication/?VW=T&DM=SLNL&PA=37325&D1=a&D2=0&D3=0&D4=0&D5=0-4,137,152,220,237&D6=0,4,9,(l-1),l&HD=130605-0936&HDR=G2,G1,G3,T&STB=G4,G5 (accessed December 2013).

Couper, P. 2005. "Technology Trends in Survey Data Collection." *Social Science Computer Review* 23: 486–501. DOI: http://dx.doi.org/10.1177/0894439305278972.

Dagevos, J. and R. Schellingerhout. 2003. "Sociaal-culturele integratie. Contacten, cultuur en orientatie op de eigen groep." In *Rapportage minderheden*, edited by J. Dagevos, M. Gijsberts, and v. C. Praag, 317–362. [In Dutch: Socio-Cultural integration. Contacts, culture and focus on the own ethnic group]. Den Haag: SCP. Available at: http://www.scp.nl/Publicaties/Alle_publicaties/Publicaties_2003/Rapportage_minderheden_2003 (accessed December 2012).

De Leeuw, E.D. 2005. "To Mix or not to Mix Data Collection Modes in Surveys." *Journal of Official Statistics* 21: 233–255.

De Leeuw, E.D., D.A. Dillman, and J.J. Hox. 2008. "Mixed-Mode Surveys: When and Why." In *International Handbook Of Survey Methodology*, edited by E.D. de Leeuw, J.J. Hox, and D.A. Dillman, 299–316. New York: Taylor and Francis Group.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society. Series B (Methodological)* 39: 1–38.

Dillman, D.A. 2007. *Mail and Internet Surveys: The Tailored Design Method* (2nd ed.). New York: John Wiley & Sons Inc.

Dillman, D.A. and L.M. Christian. 2005. "Survey Mode as a Source of Instability in Responses Across Surveys." *Field Methods* 17: 30–52. DOI: http://dx.doi.org/10.1177/1525822X04269550.

Feskens, R.C.W. 2009. *Difficult Groups in Survey Research and the Development of Tailor-made Approach Strategies*. Den Haag/Utrecht: Statistics Netherlands and

Utrecht University. Available at: http://www.cbs.nl/NR/rdonlyres/8F317AA9-1074-4BF7-84EB-015D013DBB80/0/2009x11feskenspub.pdf (accessed October 2011).

Feskens, R.C.W., J. Kappelhof, J. Dagevos, and I.A.L Stoop. 2010. *Minderheden in de mixed-mode? Een inventarisatie van voor- en nadelen van het inzetten van verschillende dataverzamelingsmethoden onder niet-westerse migranten*. SCP-special 57. [In Dutch: Ethnic minorities in the mixed mode? An inventory of the advantages and disadvantages of employing different data collection methods among non-Western migrant] Den Haag: SCP. Available at: http://www.scp.nl/Publicaties/Alle_publicaties/Publicaties_2010/Minderheden_in_de_mixed_mode (accessed December 2014).

Gijsberts, M. and J. Iedema. 2011. "Opleidingsniveau van niet-schoolgaanden en leerprestaties in het basisonderwijs." In *Jaarrapport Integratie 2011*, edited by M. Gijsberts, W. Huijnk, and J. Dagevos, 76–99. [In Dutch: Education level of persons who do not attend school and educational attainment in primary school]. Den Haag: SCP. (2012-3). Available at: http://www.scp.nl/Publicaties/Alle_publicaties/Publicaties_2012/Jaarrapport_integratie_2011 (accessed December 2014).

Graham, J.W., A.E. Olchowski, and T.D. Gilreath. 2007. "How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory." *Prevention Science* 8: 206–213. DOI: http://dx.doi.org/10.1007/s11121-007-0070-9.

Groves, R.M. 1989. *Survey Costs and Survey Errors*. New York: Wiley.

Groves, R.M. and M.P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: Wiley.

Groves, R.M. and E. Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias." *Public Opinion Quarterly* 72: 167–189. DOI: http://dx.doi.org/10.1093/poq/nfn011.

Kappelhof, J.W.S. 2010. *Op maat gemaakt? Een evaluatie van enkele respons verbeterende maatregelen onder Nederlanders van niet-westerse afkomst.* [In Dutch: An evaluation of several response enhancing measures among Dutch of non-Western origin]. Den Haag: SCP. (SCP-special 53). Available at: http://www.scp.nl/Publicaties/Alle_publicaties/Publicaties_2010/Op_maat_gemaakt (accessed December 2014).

Kappelhof, J.W.S. 2014. "The Effect of Different Survey Designs on Nonresponse in Surveys Among Non-Western Minorities in The Netherlands." *Survey Research Methods* 8: 81–98. Available at: http://www.surveymethods.org (accessed December 2014).

Kemper, F. 1998. "Gezocht: Marokkanen. Methodische problemen bij het verwerven en interviewen van allochtone respondenten." [In Dutch: Wanted: Moroccans. Methodological problems with obtaining response and interviewing respondents of foreign origin]. *Migrantenstudies* 1: 43–57.

Korte, K. and J. Dagevos. 2011. *Survey Integratie Minderheden 2011. Verantwoording van de opzet en uitvoering van een survey onder Turkse, Marokkaanse, Surinaamse en Antilliaanse Nederlanders en een autochtone vergelijkingsgroep*. [In Dutch: Survey on the Integration of Minorities 2011. Report on the design and implementation of the survey among Turkish, Moroccan, Surinamese, Antillean Dutch and a mainstream Dutch reference group]. Den Haag: SCP.

Kreuter, F. 2013. *Improving Surveys with Paradata. Analytic Uses of Process Information*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Little, R.J. and D.B. Rubin. 2002. *Statistical Analysis With Missing Data*. New York: John Wiley & Sons.

Maitland, A., C. Casas-Cordero, and F. Kreuter. 2009. "An Evaluation of Nonresponse Bias Using Paradata from a Health Survey." In: Proceedings of the Section on Government Statistics: American Statistical Association, Joint Statistical Meetings, 2009. 370–378. Alexandria, VA: American Statistical Association. Available at: http://www.amstat.org/sections/SRMS/proceedings/y2009/Files/303004.pdf (accessed December 2014).

Moorman, P.G., B. Newman, R.C. Millikan, C.K.J. Tse, and D.P. Sandler. 1999. "Participation rates in a Case-control Study: The Impact of Age, Race, and Race of Interviewer." *Annals of epidemiology* 9: 188–195. DOI: http://dx.doi.org/10.1016/S1047-2797(98)00057-X.

Reep, C. 2003. *Moeilijk Waarneembare Groepen. Een inventarisatie.* [In Dutch: Hard to survey populations: An inventory]. Voorburg/Heerlen: CBS. (H1568-03-s00).

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Schmeets, H. 2005. "De leefsituatie van allochtonen." In *Enquêteonderzoek onder allochtonen. Problemen en oplossingen*, edited by H. Schmeets en and R. van der Bie, 169–176. [In Dutch: The living conditions of Dutch of foreign origin]. Voorburg/Heerlen: CBS.

Schothorst, Y. 2002. "Onderzoek onder allochtonen: wat mag, wat moet en wat kan?" In: *Interviewen in de multiculturele samenleving. Problemen en oplossingen*, edited by H. Houtkoop en and J. Veenman, 101–116. [In Dutch: Survey research among Dutch of foreign origin: What may be done, what has to be done and what is possible?]. Assen: Koninklijke Van Gorcum.

Schouten, B. and F. Cobben. 2007. *R-Indexes for the Comparison of Different Fieldwork Strategies and Data Collection Modes*. Voorburg/Heerlen: CBS. (Discussion Paper 07002). Available at: http://www.risq-project.eu/papers/schouten-cobben-2007-a.pdf (accessed December 2013).

Schouten, B.F. and F. Cobben. 2008. *An Empirical Validation of R-Indicators*. Voorburg/Heerlen: CBS. (Discussion Paper 08006) Available at: http://www.risq-project.eu/papers/cobben-schouten-2008-a.pdf (accessed December 2013).

Schouten, B. F. Cobben, and J. Bethlehem. 2009. "Indicators of Representativeness of Survey Nonresponse." *Survey Methodology* 35: 101–113. Available at: http://www.risq-project.eu/papers/schouten-cobben-bethlehem-2009.pdf (accessed December 2013).

Schouten, B. J. Bethlehem, K. Beullens, Ø. Kleven, G. Loosveldt, A. Luiten, K. Rutar, N. Shlomo, and C. Skinner. 2012. "Evaluating, Comparing, Monitoring, and Improving Representativity of Survey Response Through R-Indicators and Partial R-Indicators." *International Statistical Review* 80: 382–399. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.2012.00189.x/abstract (accessed December 2013).

Schouten, B., N. Shlomo, and C. Skinner. 2011. "Indicators for Monitoring and Improving Representativity of Response." *Journal of Official Statistics* 27: 231–253.

Shlomo, N., C. Skinner, B. Schouten, T. Carolina, and M. Morren. 2009. *Partial Indicators for Representative Response*. Rep. No. RISQ deliverable 4. version 2. Available at: http://www.risq-project.eu/papers/RISQ-Deliverable-4-V2.pdf (accessed December 2013).

Singer, E. 2002. "The Use of Incentives to Reduce Nonresponse in Household Surveys." In *Survey Nonresponse*, edited by R.M. Groves, D. Dillman, J.L. Eltinge, and R.J.A. Little, 163–177. New York: John Wiley & Sons.

Singer, E., J. Van Hoewyk, N. Gebler, T. Raghunathan, and K. McGonagle. 1999. "The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys." *Journal of Official Statistics* 15: 217–230.

Singer, E., J. Van Hoewyk, and M.P. Maher. 2000. "Experiments With Incentives in Telephone Surveys." *Public Opinion Quarterly* 64: 171–188. DOI: http://dx.doi.org/10.1086/317761.

Smulders, M. 2011. *Onderzoeksverantwoording Survey Integratie Minderheden 2011* [In Dutch: research description on the Survey on the Integration of Minorities 2011]. Dongen: GFK Panel Services Benelux B.V.

Stoop, I.A.L. 2005. *The Hunt for the Last Respondent: Nonresponse in Sample Surveys*. Den Haag: SCP.

Särndal, C.-E. 2011. "Three Factors to Signal Non Response Bias With Applications to Categorical Auxiliary Variables." *International Statistical Review* 79: 233–254. DOI: http://dx.doi.org/10.1111/j.1751-5823.2011.00142.x.

Särndal, C.-E., and S. Lundström. 2010. "Design for Estimation: Identifying Auxiliary Vectors to Reduce Nonresponse Bias." *Survey Methodology* 36: 131–144.

Van Ingen, E., J. De Haan, and M. Duimel. 2007. *Achterstand en afstand. Digitale vaardigheden van lager opgeleiden, ouderen, allochtonen en inactieven*. [In Dutch: Lagging behind. Digital skills of lower educated, elderly, foreign origin and inactive persons]. Den Haag: SCP. (SCP- 2007/24).

Veenman, J. 2002. "Interviewen in multicultureel Nederland." In *Interviewen in de multiculturele samenleving. Problemen en oplossingen*, edited by H. Houtkoop en and J. Veenman, 1–19. [In Dutch: Interviewing in the multi-cultural Netherlands]. Assen: Koninklijke Van Gorcum.

Voogt, R.J.J. and W.E. Saris. 2005. "Mixed Mode Designs: Finding the Balance Between Nonresponse Bias and Mode Effects." *Journal of Official Statistics* 21: 367–387.

Wagner, J. 2008. *Adaptive Survey Design to Reduce Nonresponse Bias*, Michigan: University of Michigan. Available at: http://deepblue.lib.umich.edu/bitstream/2027.42/60831/1/jameswag_1.pdf (accessed December 2014).

Wagner, J. 2010. "The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data." *Public Opinion Quarterly* 74: 223–243. DOI: http://dx.doi.org/10.1093/poq/nfq007.

Received March 2013
Revised December 2014
Accepted December 2014