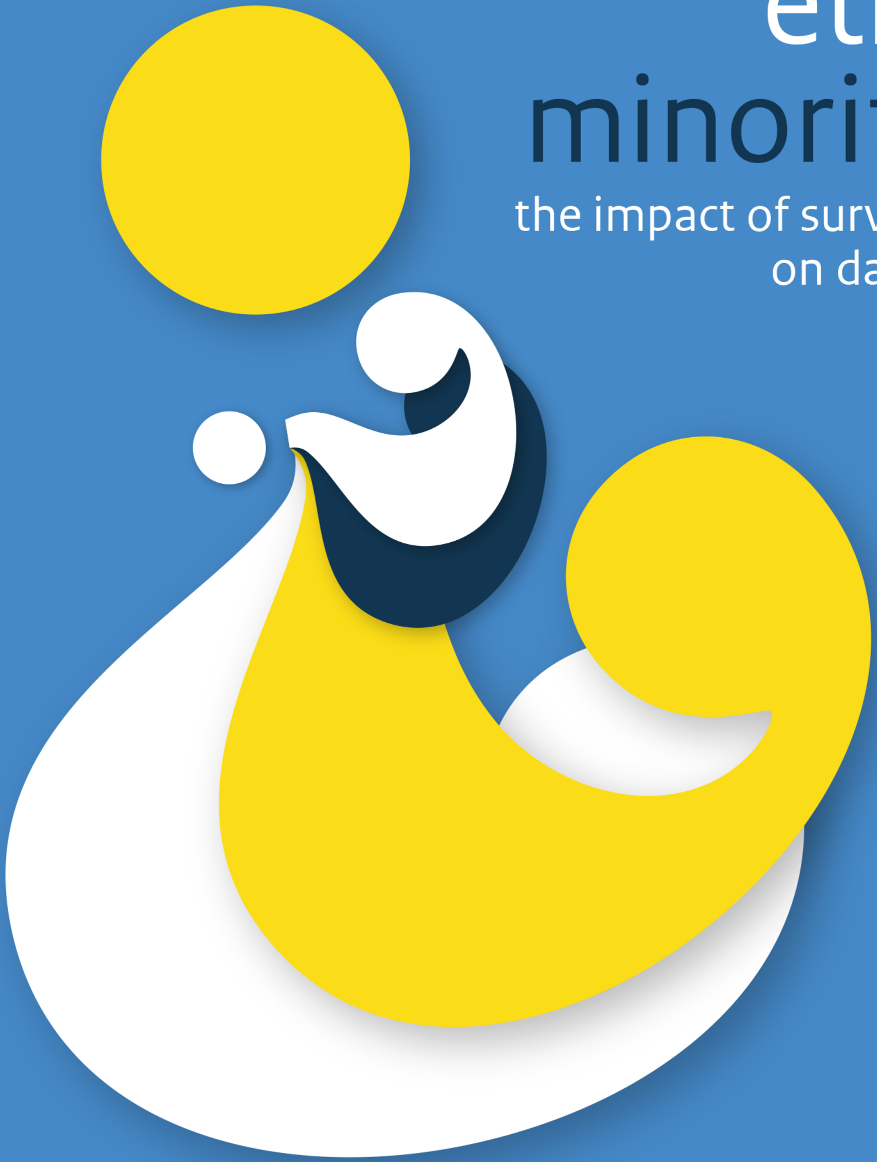
Surveying ethnic minorities:

the impact of survey design on data quality



Joost Kappelhof

Surveying ethnic minorities: the impact of survey design on data quality

ISBN 978 90 377 0739 7 (dissertation)
ISBN 978 90 377 0545 4 (trade paperback)

Layout and design: Textcetera, The Hague
Cover design: Groeier!, Zwaag

Surveying ethnic minorities: the impact of survey design on data quality

Enquêteren onder etnische minderheden:
de invloed van het enquête-ontwerp op data kwaliteit

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht op gezag van de
rector magnificus, prof. dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor
promoties in het openbaar te verdedigen op
vrijdag 19 juni 2015
des middags te 12.45 uur

door

Johannes Willem Simon Kappelhof

geboren op 27 augustus 1974
te Hoorn

Promotor:	Prof. dr. E.D. de Leeuw
Copromotor:	Dr. I.A.L. Stoop

Beoordelingscommissie: Prof. dr. J.W.M. Das (Tilburg University)
Prof. dr. P.A. Dykstra (Erasmus University Rotterdam)
Prof. dr. ir. A.G. van der Lippe (Utrecht University)
Prof. dr. G. Loosveldt (K.U. Leuven)
Prof. dr. J.J.G. Schmeets (Maastricht University)

Acknowledgements

I would like to take this opportunity to express my gratitude and appreciation for several individuals that have made the successful completion of this dissertation possible. First of all, I wish to thank my promotor Edith de Leeuw and my co-promotor Ineke Stoop. They have been instrumental in making this project a success. I would like to thank them both for giving me the opportunity to learn so much from them. I am grateful for and highly appreciative of Edith's support, input, goal-oriented guidance and critical, but always relevant comments. I would like to thank Ineke for the support, invaluable input and comments, but also for stimulating me to pursue a Ph.D. You are not only a highly valued colleague, but a friend and I am a better person for knowing you.
I would also like to thank the Netherlands Institute for Social Research/SCP for creating the possibility for me to work on my Ph.D. Furthermore, I thank my colleagues at SCP for their support and their interest in the progression of this Ph.D. I would also like to thank some of the colleagues at SCP individually. First of all, I would like to thank Jaco Dagevos who has not only given me important advice and insightful comments during this project, but, as head of the research section on Education, Minorities and Methodology, also saw to it that I would have the time to write up my thesis during these last two and a half years. It has been greatly appreciated. Secondly, I would like to thank my former I&A colleagues: Patty Adelaar, Erminio Ballerini, Jeroen Boelhouwer, Leen Both, Henk Fernee, Sarah Kanneworff and Jan Spit for being supportive and always interested in my progress. I especially thank Jurjen Iedema whose willingness to listen to my ideas and provide me with feedback has helped me immensely.
I would also like to thank the anonymous reviewers for the useful, but critical comments. Their comments helped improve chapters 3, 4, 5 and 6, which, as a result, have been published or accepted accepted for publishing.
I would also like to express my gratitude to my family and friends, whose support and interest in my dissertation I have always greatly appreciated. Furthermore, they have always shown me the value of spending time with loved ones. And last, but not least I would like to thank my fantastic wife Irina and my beautiful children Maria Sara and Hanna Louise, whose love and support helped me during challenging times. Furthermore, I am not sure if I'll ever find the right words to show my deep gratitude and appreciation for all the work my wife Irina has done on proofreading and correcting my English texts. I would like to dedicate this dissertation to you.

Joost Kappelhof
Den Haag, February 2015

Table of Contents

1 Introduction

In the Netherlands, the policy on minorities dates back to the 1980s. Before the 1980s, the idea had been that *guest labourers*, as they were called back then, would return to their country of origin and therefore a specific integration policy was not necessary (Scholten 2011; WRR 1989). It is no wonder that research on the living conditions of minorities only started in the mid-1980s. At that time a policy on minorities started being developed and monitoring was deemed necessary, particularly to tackle the perceived disadvantaged socio economic position of minorities. Furthermore, the first publications on minorities' policy in the 1980s showed that there was a lack of reliable data on migrant groups (WRR 1989, p. 183). As a result, it was not always possible to make claims or give accurate estimates on the socioeconomic position of minorities.
Nowadays, more information is collected and recorded in the Netherlands about the socioeconomic position and sociocultural integration of minorities. A large portion of this information is to this day still collected by means of surveys and, even though the amount of data has increased, the question about its accuracy remains very relevant.
In its 1989 report, the Scientific Council for Governmental Policies (WRR) uses the terms *"non-autochthonous"* (in Dutch: *allochtoon*) and *minority* to describe the group being researched. The WRR report (1989, p. 14) states: "In this report the Council (WRR) interprets *non-autochthonous* as: not of Dutch descent. *Non-autochthonous residents* are foreigners in the legal sense, ex-foreigners who have acquired the Dutch nationality, Dutch who come from former colonies, and their descendants until the third generation, as long as they view themselves as foreign. An *ethnic minority* is defined as a group of foreigners in a disadvantaged [socioeconomic] position."
In the last few decades, the policy on minorities has mainly focused on the position of non-Western minorities[1], although nowadays there is also a noticeable increase in attention for the more recent minority groups, originating from Central and Eastern European countries. In 2014, non-Western minorities made up about 12% of the population in the Netherlands (CBS-statline).

In the present thesis we set out to investigate the quality of survey data collected among non-Western minorities in the Netherlands and how this might relate to the survey design. We focus on the two quality dimensions that seem most pertinent for data about non-Western minorities in the Netherlands: accuracy and comparability.
We pay particular attention to two aspects of accuracy: 1) representation, that is, how well the population is reflected by the respondents to the survey, and 2) measurement, that is, to what degree the manner of administering the survey allows for an accurate measurement of the substantive topics. With regard to comparability, we focus on how

[1] Statistics Netherlands uses the following official definition to describe a non-Western person in the Netherlands: "Every person residing in the Netherlands with at least one parent born in Africa, Latin-America, Asia (excluding Indonesia and Japan) or Turkey (Reep 2003).

comparable the survey data collected between different minority groups are. The main outline of the chapters is as follows:

Chapter 2 provides the theoretical framework. It looks at the difficulties concerning the definition of ethnicity and ethnic minorities, and their consequences. It also provides an overview of the literature concerning the problems that can arise when conducting surveys among ethnic minorities. These problems are then correlated to specific *Total Survey Error* sources (Biemer and Lyberg 2003, Biemer 2010; Groves 1989; Groves et al. 2009). The TSE-concept is a theoretical framework of all possible types of error that can arise during the development and implementation of the survey, as well as during the collection and processing of the survey data.
Furthermore, chapter 2 provides an overview of measures designed to increase response rates among minority groups, such as translated questionnaires or the use of different modes of data collection, and it discusses ways of assessing the success of such measures. Attention is also paid to the trade-off between representation and measurement; in other words to what degree might the measures taken to ensure a better representation of the population affect the measurement of substantive variables among that population. The last section of chapter 2 approaches the issues of comparability and timeliness of data collected among ethnic minorities. The focus is on the ways in which survey design choices may negatively influence quality in terms of comparability and timeliness of the data collected among ethnic minorities. Finally, cost-related considerations are presented in connection to different survey designs and it is discussed how they should be included in the trade-off between quality and cost of data collection among ethnic minorities.
In surveys, the final respondents or the achieved sample are expected to represent the target population. Chapter 3 describes a quasi-experimental study based on eight sub-surveys conducted among non-Western minorities in the Netherlands. The aim of the study is to find out how survey design choices – such as the use of bilingual interviewers with a shared ethnic background or the use of a reissue phase – affect the representativity and the potential for nonresponse bias on survey estimates.
Chapter 4 describes an experimental study that investigates how the use of different methods of data collection among minority groups affects representativeness. To this end, the quality of the achieved samples based on a single-mode CAPI survey design is compared with the quality of the achieved samples based on a sequential mixed-mode survey design among four non-Western minority ethnic groups in the Netherlands.
In surveys, the way a question is asked or presented can affect the respondents' answers. Chapter 5 describes an experimental study that investigates to what degree the measurement of survey questions among non-Western minorities in the Netherlands is affected by different aspects of the survey design: the use of different data collections modes, a translated questionnaire, conducting an interview in the native language or using an interviewer with a shared ethnic background.

The comparability of survey estimates between subgroups or different surveys can be threatened by a multitude of factors. Chapter 6 describes an experimental study that intends to find out to what degree method bias (i.e., unwanted systematic methodological impact on the measurement) introduced as a result of interviewer effects (such as the ethnicity and the gender of the interviewer, but also interview language), the presence of influential others during the interview, and differences in the sociodemographic composition between samples may affect the comparability of survey estimates among four non-Western minority groups in the Netherlands.
Chapter 7 provides a summary of the main findings of the previous chapters. It will then apply these results to inform on the main research topic of this study: the quality of survey data among non-Western minorities in the Netherlands and how this might relate to survey design, and will discuss how these results may be of use to a wider audience.

2 Survey research and the quality of survey data among ethnic minorities

This chapter discusses four key topics in designing and evaluating survey research among ethnic minorities for policy makers. First of all, it discusses the difficulties concerning the use of the terms ethnicity and ethnic minorities. Secondly, it reviews the challenges and pitfall as to why ethnic minorities are difficult to survey. To this end, an overview of the international empirical literature on reasons why it is difficult to conduct survey research among ethnic minorities will be placed in the TSE framework. Thirdly, it discusses measures that can be undertaken to increase the representation of minorities in surveys and it discusses the consequences of these measures. In particular the relationship with survey design, sample frame and trade-off decisions in the TSE paradigm is discussed in combination with budget and time considerations. It also reviews the empirical literature on different methods that can be applied to assess the data quality of surveys among ethnic minority groups and how this can be utilized to assess the representation and measurement of ethnic minorities in national surveys. The fourth part deals with potential sources of method bias that can arise as a result of survey design choices in surveys among ethnic minority groups. In particular how this can affect the cross cultural comparison of survey results when surveying different ethnic minority groups. This part will draw on existing international literature about cross-cultural survey research and best practices.[1]

2.1 Introduction

There are important reasons for collecting information about ethnic minorities. Many national governments are, for instance, interested in the degree of sociocultural integration of ethnic minorities, but also in the developments regarding their socioeconomic position (Bijl and Verweij 2012; Font and Mendez 2013; Thomas 2008). In turn, the healthcare sector poses important questions about possible differences in lifetime risk of mood, anxiety and substance use disorders (Breslau et al. 2005), addiction-related behaviour (Caetano et al. 1998) or use of facilities (Herníndez-Quevedo and Jiminez-Rubio 2009; Wen et al. 1996). In the face of lack of data in some countries of origin, research among recent migrants can also provide estimates of the rates of incidence of certain diseases (Chaturvedi and McKeigue 1994). To this effect, survey research remains an important method of obtaining information about these special groups.
Setting up and conducting survey research is not easy (e.g., low response rates, measurement problems, coverage errors, etc.). In survey research among ethnic minorities,

[1] An abbreviated version of this chapter has been accepted for publication as Kappelhof, J.W.S. (2016). Ethnic minorities in surveys: applying the TSE paradigm to surveys among ethnic minority groups to assess the relationship between survey design, sample frame and survey data quality. In P. Biemer, E.D. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, C. Tucker, and B.T. West (Eds.), Total Survey Error in Pratice (Chapter 16). Hoboken, New Jersey: John Wiley and Sons.

these problems are often amplified (Font and Mendez 2013). This can lead to underrepresentation of ethnic minorities in surveys targeting the general population or to surveys focused on ethnic minorities delivering an incomplete picture of their target population (Feskens et al. 2007). In both scenarios, the data will be insufficient for an accurate assessment of the position of ethnic minorities in connection to the topic of interest.

The Total Survey Error (TSE) framework offers a way of looking at the quality of survey data (Groves 1989; Groves et al. 2009; Groves and Lyberg 2010; Biemer and Lyberg 2003). Biemer (2010, p. 817), describes TSE as "[...] the accumulation of all errors that may arise in the design, collection, processing, and analysis of survey data. In this context, survey error is defined as the deviation of a survey response from its underlying true value."

The TSE framework takes into account both the measurement (e.g., construct validity, measurement and processing error) and the representation (e.g., coverage, sampling, non-response and adjustment error) of the target population (Groves and Lyberg 2010).

The present chapter does not presume to offer a full account of TSE, its main objective being to make researchers, policy makers, etc. aware of the difficulties related to survey research among ethnic minorities. TSE provides a useful framework for placing the difficulties, but also the possible choices connected with survey research among ethnic minorities and consequences thereof, in the appropriate context.

The TSE approach focuses primarily on the accuracy of the data, described in terms of variance and bias. The smaller the variance (random error) and the smaller the (squared systematic error) bias, the more accurate and, consequently, of better quality the estimate is considered to be. However, to assess the quality of surveys and survey data among ethnic minorities, it is also interesting to consider other quality dimensions (see, for instance, Eurostat 2000 or OECD 2011). Among these, two dimensions are of special interest. The first is comparability which refers to the impact of differences in applied statistical concepts and definitions, but also in question wording or data collection methods when statistics are compared between geographical areas, non-geographical domains (groups) or reference periods. The second is timeliness which refers to the speed with which data becomes available. In fact, with surveys among ethnic minorities data often needs to be available rather quickly. For example, when an ethnic minority group suddenly becomes politically relevant and the government needs to inform the political debate. Furthermore, it is frequently compared to data on the general population, data from other sources and/or from other reference periods, which may have been collected in a different way. At the same time, data quality is never considered independently. Factors such as funding and time place limitations on the maximum level of quality that can be obtained.

The present chapter has the following structure: the next section goes into the difficulties concerning the definition of ethnicity and ethnic minorities and their consequences. Section 2.3 describes the problems that can arise when conducting surveys among ethnic minorities. These problems are then linked to specific TSE error sources within the representation and the measurement dimensions. The following section describes the measures that can be taken in order to ensure a better representation of ethnic minorities in surveys and ways to assess the success of such measures. Attention is also being paid to the trade-off: to what degree might the measures taken to ensure a better representation of the population affect the measurement among that population. Several procedures to minimize measurement error as a result of these response-enhancing measures will be mentioned as well. The last section of this chapter approaches the issues of comparability and timeliness of data collected among ethnic minorities. We focus on how survey design choices may negatively influence quality in terms of comparability and timeliness of the data collected among ethnic minorities. Finally, we take a look at cost-related considerations in connection to choices made in the design of surveys to the effect of improving the quality of data collected among ethnic minorities.

2.2 On the use of the terms *ethnicity* and *ethnic minorities*

Ethnicity is often used as an important explanatory variable. Nevertheless, defining ethnicity and classifying persons on the grounds of ethnic status is not easy and it represents an aspect deserving particular attention, among others, in policy and official statistics (Simon 2007), sociologic and migration studies (Aspinall 2002; Jacobs et al. 2009) and in health studies (Bhopal 2004; Chaturvedi and McKeigue 1994).

A first problem in defining ethnicity lies in the subjective, multi-faceted and changing nature of ethnic identification and the lack of consensus on what an “ethnic group” is (Jacobs et al. 2009, p. 15). In this respect, an important thing to bear in mind is that the concept of ethnicity should rather be seen as a process than as a clear-cut determination. This becomes obvious when asking questions such as: “How far back in time should one go in order to determine ethnicity?” or “In which context does ethnicity get defined: which groups are being distinguished, under what circumstances and on what grounds?”

A second problem resides in measuring the concept of ethnicity. A multitude of „ethnic” indicators (e.g., race, country of birth, nationality, citizenship) are being used to determine ethnicity (see, for instance, Simon 2007). Moreover, these indicators cover the concept „ethnicity” only in part, each having a unique component that does not overlap with the others (Chaturvedi and McKeigue 1994; Erens 2013; Ozaki and Sue 1995).

Simon (2007) reviews in this sense three classification approaches that countries traditionally adopted for measuring ethnicity. The first is a state-centred model, in which country of birth and citizenship variables are collected. This approach is more common in countries where immigration is a newer phenomenon. The second is a mosaic model in which variables on nationality/ethnicity and language are collected. This approach is more frequently used in countries with autochthonous minorities, like, for instance, the Roma in Romania. Or in countries where the autochthonous population is a minority such as native Americans and Inuit in Canada. The third possibility is of the “post-migration multicultural type” in which info on ethnic group, religion and/or parents’ country of birth is gathered. This happens in countries with a tradition of immigration, where ethnicity is substituted with migrant origin, so as to include not only recent immigrants, but also their descendants (Font and Mendez 2013).

The different traditions and indicators being collected have consequences for the survey sponsor, the researcher and the user of ethnicity data. Not only is it difficult to classify ethnicity in, for instance, statistical categories, but also comparisons between studies, groups, or countries are not at all straightforward (see Feskens et al. 2006). In the case of comparative research therefore it is essential to take into account the operationalization of the term, so as to ensure comparability. This is also the case for interpreting research results.

There is one more observation to be made about the sensitivity of ethnicity data and particularly of data on ethnic minorities. The sense in which „minority" is often used to encompasses not only a numerical minority, but also the presupposition that the group as a whole holds a socioeconomically or politically disfavoured or non-dominant position (Jacobs et al. 2009). Caution should at all times be exerted in the use, production and publication of ethnicity data. As Simon (2007) states on page 15, "if racial or ethnic minority stereotypes are the product of racism, then the use of ethnic of racial categories is sure to affirm them".

2.3 On the representation and measurement of ethnic minorities in surveys

Difficulties regarding the representation and the measurement of minorities arise both in surveys targeting ethnic minorities and in surveys of the general population in which ethnic minorities are represented. The overview provided here is not meant to be exhaustive, but should give the reader a clear picture of the main problem areas and of the multitude of factors that can affect survey outcomes among ethnic minorities and how they relate to the TSE framework.

2.3.1 Sampling frame, sampling, coverage and coverage related issues among ethnic minorities

When a general population survey is being done in which ethnic minorities are represented it may be that the type of sampling frame rather than the lack of a usable sampling frame should lead to a more serious undercoverage of ethnic minorities (Chaturvedi and McKeigue 1994; Mendez and Font 2013). For example, the use of a national register as a sampling frame will often lead to the undercoverage of ethnic minorities, since recent immigrants and illegals are not registered (yet) and ethnic minorities are overrepresented in those categories (Dourleijn 2010; Rinken 2013; Weltevree et al. 2009). At the same time, mode-dependent sampling frames can lead to a higher degree of undercoverage and thus exclusion of ethnic minorities. For example, the use of a telephone-based sampling frame excludes ethnic minorities to a greater degree because ethnic minorities less often own a phone or a fixed landline (Lipps and Kissau 2012; Schothorst 2002; Thornberry and Massey 1988). Also, the slow registration or lack of timely updates of sampling frames has a larger impact on ethnic minorities. They appear to be more mobile, which in turn leads to higher numbers of 'wrong' addresses among sampled ethnic minority members (Chaturvedi and McKeigue 1994; Rinken 2013). Technically speaking, this is nonresponse. However, the inability to locate sampled individuals at the sampled address and the lack of information about a new address can also mean that the sampled individual is ineligible if they no longer were part of the target population at the time the sample was drawn (e.g., moved abroad, which is not uncommon for recent immigrants), but the sampling frame was slow to update. In this case it would be a coverage error.

In case of surveys targeting exclusively ethnic minorities, several additional representation problems may appear. The biggest problem is the lack of direct or usable ethnic identifiers in a sampling frame. A list of all inhabitants of a certain country may be available, but there is no direct information at hand about the ethnic identity of individuals (Chaturvedi and McKeigue 1994; Dutwin and Lopez 2014; Erens 2013). Sometimes there are indirect identifiers, but the use thereof often leads to serious undercoverage/exclusion of ethnic minorities. For instance, information about a person's country of birth may be available. However, this identifier excludes the descendants of immigrants, who are often considered part of the target group. Another commonly used ethnic identifier is the surname, but this can also easily lead to exclusion of persons belonging to the ethnic group. Persons married outside the group who have taken over the surname of their partners and persons with a surname common in both the ethnic group and in the general population would be in this situation. More recently screening methods have been developed that use more than the surname (see for example, Mateos et al. 2007 or Schnell et al. 2013), but they are not equally efficient among each ethnic minorities population.

The lack of useful individual level ethnic identifiers in the sampling frame can also lead to the decision to exclude members of the ethnic target population beforehand due to cost-efficiency reasons (Dutwin and Lopez 2014; Erens 2013). For instance, a choice could be made to exclude areas in which it is believed very few members of the ethnic minority live. A related design decision affecting coverage happens when surveys designed to target several ethnic minority groups leave out the 'smaller' ethnic minority groups from the target population. One reason for this choice could be that they are geographically too dispersed and small in numbers so it would be too costly to conduct a survey among them (Erens 2013).

Another coverage related issue has to do with the mismatch between the social reality of the ethnic minorities living conditions and part of the commonly used target population definition "living in a private household". It is not uncommon for certain ethnic minority groups, such as migrant labourers, to live in communal or collective households (Barnes 2008).

Finally, one specific sampling issue needs mentioning. Within certain ethnic minority groups it can happen more often that several families live at the same address (Duque et al. 2013; Schmeets 2005). In the case of address-based sampling, this can lead to bias in the selection of the correct family and/or sampled person when constructing a sample consisting of a single person at each address.

2.3.2 Factors affecting nonresponse among ethnic minorities

There are three main reasons for sampled persons' failure to participate (Groves et al. 2009): the inability to locate or *contact* the sampled person, the *refusal* of the sampled person to participate, and the *inability* of the sampled persons to participate. Groves

& Couper (1998) distinguish a number of general factors that influence both cooperation and the likelihood of contact. These are *societal environmental attributes*, *sociodemographic attributes of the sampled person*, *survey design features* and, in case of interviewer-assisted mode, also *the interviewer* and *the interaction between the interviewer and the sampled person* (plus the interaction with the person answering the phone or opening the door). Several of these factors can also affect the ability of the sample unit to participate. Furthermore, *physical impediments* and *accessible at-home patterns* are also distinguished as factors influencing the likelihood of contact in case of in-person interviewer-assisted surveys (Groves and Couper 1998). There is a large amount of literature available about causes for nonresponse in general; therefore we shall only discuss specific attributes within these general factors that elicit nonresponse to a greater degree among ethnic minorities compared to the general population.

Several *environment* attributes on a *societal* level (i.e., global characteristics affecting the survey taking climate) and on a more local level (i.e., neighbourhood characteristics) can be found that will increase the probability of nonresponse among ethnic minorities more than in the general population (Barnes 2008; Groves and Couper 1998; Deding et al. 2008; Erens 2013; Feskens et al. 2007; Laganá et al. 2013; Lipps et al. 2013; Mendez and Font 2013; Thomas 2008). First of all, it is more common among ethnic minorities to feel excluded to some extent (e.g., do not feel recognized by society, are unfamiliar with the customs of the host country, feel outside the mainstream society or have different views as to what are important research topics) which causes them to be less inclined to participate in surveys. The less an ethnic minority group identifies with the core values and beliefs of the host or mainstream society in case of an indigenous minority or the more distant it feels in social, linguistic and cultural terms, the more likely it is not to find the survey topics equally important – which, in turn, will result in lower response rates. Secondly, lower levels of trust and questioning the legitimacy of the survey organisation are more common among ethnic minorities, especially among more recent groups originating from countries without a functioning democracy. Thirdly, a negative perception of the ethnic minority group by the host country or mainstream society can affect response rates, especially if its members often encounter racist attitudes. A fourth societal level explanation is survey fatigue or being over-surveyed. However, survey fatigue is less likely to be a reason for underrepresentation if the ethnic minority member has only lived in the host country for a short while as is the case with recent immigrants.

Highly urban areas with a high number of flats or apartment blocks with a central locked entrance can represent a local level factor. Ethnic minorities are overrepresented in highly urban areas (Deding et al. 2008; Feskens et al. 2007; Groves and Couper 1998). Furthermore, in certain countries they are more likely to reside in households with *physical entry barriers* (Groves and Couper, 1998). Big city dwellers and households with physical entry barriers have been identified as having higher non-contact and refusal rates than other dwelling types (Groves and Couper 1998; Stoop 2005).

The *sociodemographic characteristics of sampled persons* that affect nonresponse have been studied quite extensively (see for example, Bethlehem et al. 2011; Groves and Couper 1998; Stoop 2005). A comparison – based on various studies – between ethnic minority groups and the native or majority population reveals that ethnic minority groups are on average younger and are overrepresented in lower socioeconomic positions. That is, they tend to have lower employment rates – especially women – and job stability, and are more often single parents, are overrepresented in jobs with non-standard work times and generally have lower educational levels (Deding et al. 2013; Feskens et al. 2007; Gijsberts and Iedema 2011; Laganá et al. 2013; Morales and Ros 2013; Smith 2013). These sociodemographic characteristics have all been shown to negatively impact response rates via lower contact rates and/or lower cooperation rates (Groves and Couper 1998; Feskens et al. 2007; Stoop 2005). Furthermore, the status of the sampled person in terms of regularity of their stay in the country greatly affects the likelihood of survey participation and ethnic minority migrant labourers are overrepresented in the category of irregular migrants (Barnes 2008).

The underrepresentation of ethnic minorities in surveys can be partly explained by the 'standard' set of *survey design features*, the implementation and the survey culture within a country. There is a wide variety of implicit or country specific assumptions that are made about survey processes, such as what is considered a standard household, daily routines, legitimate survey requests and incentives (Laganá et al. 2013). However, as Laganá et al. (2013) point out, these implicit assumptions about what is standard derive from the experiences of survey researchers and interviewers and these experiences are usually based on surveying the majority population. In the survey design it is often forgotten that the 'standard' set of *survey design features* and call strategies does not necessarily correspond with the social reality of the ethnic minorities. Similarly, certain *survey design choices* can influence the ease with which ethnic minority members are able to participate. Below we provide a list of examples as to how certain, more 'standard' *survey design features* can actually contribute to increased nonresponse among ethnic minority members.

Firstly, and most obviously, it is well documented that difficulties in understanding the main or national language among ethnic minority groups can contribute to higher levels of nonresponse, as the sampled persons are not able to understand the survey request and survey questions (Barnes 2008; Deding et al. 2008; Dutwin and Lopez 2014; Feskens et al. 2006; Smith 2013).

Secondly, ethnic minority groups – especially young men in these groups, ethnic minorities' inner city dwellers and recent immigrants – have a higher residential mobility (Barnes 2008; Morales and Ros 2013; Mendez and Font 2013; Mendez et al. 2013). Quite often the sampling frame does not keep up with the higher residential mobility. This results in higher non-contact rates in individual based samples, because there is no forwarding address available or no further action is taken. A special case are the illegal immigrants, who are obviously not often to be found in any register, but whose transient nature (e.g., staying very briefly at any single address) also makes them almost impossible to contact, even if one employs address-based sampling instead of individual named-based sampling (Barnes 2008).

Thirdly, the choice of a survey mode can contribute to nonresponse among ethnic minorities because of incomparably high refusal rates, lower penetration levels and/or functional illiteracy. For example, the use of the telephone can result in very high refusal rates among ethnic minorities (Korte and Dagevos 2011; Myrberg 2013; Schothorst 2002). Furthermore, the use of web or fixed landline telephone may also exclude ethnic minorities from participating to a greater degree, due to lower internet or telephone penetration rates (Van Ingen et al. 2007; Schothorst 2002; Lipps and Kissau 2012). In case of a register-based sampling frame, this type of exclusion counts as nonresponse and not as coverage error, because the survey mode prevents sampled persons from participating, while the sampling frame does not exclude them from being selected in the gross sample. Functional illiteracy can affect the suitability of self-completion online or postal questionnaires. Several studies have found higher levels of functional illiteracy especially among the elder ethnic minority members from second or third world countries (Schothorst 2002; McManus et al. 2006). Another issue affecting the suitability of interviews, both CAPI and CATI, is the fact that ethnic minorities are overrepresented among persons with less *accessible at-home patterns*. They more often have jobs with non-standard working hours (Deding et al. 2008; Feskens et al. 2006) and some ethnic minority groups may have different culturally determined at-home patterns which causes them to be found at home less often. For example, Kemper (1998) studied Moroccans living in the Netherlands and discovered that especially men are often away from home; staying outside on the street or in coffeehouses. Also, it is more common to find a man that refuses the survey request on behalf of his wife but not the other way around (Deding et al. 2008). Possibly this is partly caused by the lack of a gender match when the *interviewer interacts* with the person opening the door.

Fourthly, certain survey topics can contribute to nonresponse among ethnic minorities because they are considered sensitive or controversial (Groves and Couper 1998; Mendez and Font 2013). However, what is considered controversial or sensitive among various ethnic minority groups can vary and may not always be understood by the researcher – or simply cannot be avoided.

Last but not least, the choice of a fieldwork period and length is another example of how *survey design* choices can contribute to nonresponse among ethnic minorities. It is more common for certain ethnic minority groups, such as migrant labourers and recent immigrants, to be unavailable during the entire fieldwork period, possibly because they are visiting their native country for an extended period of time (Blohm and Diehl 2001 as cited by Feskens et al. 2006; Deding et al. 2013; Mendez et al. 2013).

2.3.3 Post survey adjustment issues related to surveys among ethnic minorities

Post hoc weighting adjustment is a frequently used method for translating the results of the sample to the entire population. A variable commonly used in the weighting model is ethnicity (Lipps et al. 2013). Often enough, the categories of the „ethnicity" variable are quite broad when it comes to general social surveys (Dutwin and Lopez 2014; Lipps et al. 2013). The choice of including a variable with such broadly defined categories for ethnic origin in the weighting model has as a consequence that the accuracy of the data about certain ethnic subgroups becomes questionable, all the more so when some ethnic subgroups are severely underrepresented in the survey.

2.3.4 Measurement issues when surveying ethnic minorities

A large body of research shows that the ethnicity of the interviewer influences the

answers of the respondent (Anderson et al. 1988; Davis 1997; Finkel et al. 1991; Van't Land 2000). This is not an effect limited to the case of ethnic minorities. A „race of interviewer" effect is also found, for instance, when white respondents are being interviewed by a black interviewer (see, for example, Hatchett and Schuman 1975). This measurement effect seems therefore more an effect of ethnic in-group versus ethnic out-group, although it is not always systematic and it seems to surface mainly on ethnic topics (Van Heelsum 2013). Moreover, the „race of interviewer" effect is generally only considered an important issue in the case of research among ethnic minorities (see section 2.4.2: on the trade-off between representation vs. measurement). For a review of the research on how the ethnicity of the interviewer can influence the answers of respondents, we refer to Van Heelsum (2013). We shall focus below on a series of causes of measurement error that play a role more often or mainly in the case of ethnic minorities.
Difficulties in understanding the interview language are more frequent among ethnic minorities (Dutwin and Lopez 2014; Feskens et al. 2006). Poor command of the original interview language can lead to measurement bias in different ways. On the one hand, poor understanding of the interview language can elicit wrong answers, while on the other hand, in the case of translated questionnaires, translation can introduce systematic differences (Harkness 2007).

The use of family members as interpreters also increases the chance of measurement differences if these persons translate and interpret the question „on the fly", or possibly even answer it for the respondent (Harkness et al. 2008). Bias can also be introduced by the use of proxy interviews (Stoop et al. 2010), for instance when the person answering the questions is not well aware of the opinion of the targeted respondent.

Differences in the interpretation of concepts being measured in surveys is a typical measurement issue commonly found when surveying people originating from different cultural backgrounds, as is the case with many ethnic minority groups (see for instance, Hui and Triandis 1989). It is also not uncommon for specific answering strategies, like extreme response styles or acquiescence, to be more typically used among specific ethnic minorities (Morren et al. 2012a; 2012b). Unfamiliarity with the nuances that distinguish the answering categories in the interview language may explain such behaviour, but also, as mentioned before, differences in the interpretation of the concept being measured.

The interview setting can, in turn, affect the answers provided by respondents, because of, for instance, the presence of influential third persons. Naturally, this effect is not specific for ethnic minorities, but it seems to surface more often in this case (Veenman 2002). A possible partial explanation might reside in linguistic problems and cultural etiquette.

2.4 On increasing participation among ethnic minorities

Groves and Couper (1998) offer an overview of survey features that fall within the span of control of the researcher and affect the contactability or the likelihood of cooperation of sampled persons in face-to-face household surveys (Table 2.1). When designing a survey, one can decide to make use of some of the features in Table 2.1 in order to increase contactability and cooperation, and subsequently the level of response.

Table 2.1
Survey design features affecting participation in face-to-face surveys (Groves and Couper 1998)

Features affecting contactability	Features affecting cooperation
– Number and schedule of calls	– (mentioning or not mentioning) the survey sponsor
– Length and timing of data collection period	– Advance letter (use of official stationary, letter signed by a person in authority, personalized, including incentive, etc.)
– Gathering information about the sample household to guide future calls	– Respondent incentives (monetary/nonmonetary, conditional/unconditional, reluctant persons)
– Interviewer workload	– Interviewer incentives
	– Respondent rules (e.g., recruitment protocols for within household selection)
	– Survey burden (length, sensitivity and cognitive burden)
	– The role of interviewers (training and briefing, highlighting positive salient aspects of the survey topic, appeal to helping behaviour)
– Follow-up procedures (different interviewer, reminders/persuasion letters, different survey mode)	

A response-enhancing measure consists of the intentional application or manipulation of such a feature to the effect of increasing survey participation. This can be done for each sampled person (e.g., an advance letter), but also just for specific subsets of sampled persons (e.g., a reissue of all or certain non-respondents, offering an incentive to reluctant sampled persons). Depending on the measures chosen, the moment of decision as to their necessity and their target, the survey design can simply be *multi-phase* or a more flexible design approach like a *responsive* survey design (Groves and Heeringa 2006) or an *adaptive* survey design (Wagner 2008; 2010).

Most features in Table 2.1 are not only applicable to face-to-face household surveys, but also to person surveys or to surveys using other data collection methods. However, sometimes the difficulty of contacting persons from a certain target population can lead to choosing a different mode than face-to-face as the main or single-mode of survey. A reason could be, for instance, that the times at which these persons are available don't overlap with the working hours of interviewers.

Furthermore, different survey modes can be used (sequentially or concurrently) to increase participation. This is known as a mixed-mode design (De Leeuw 2005). There is a range of possible designs in which different methods of data collection are used and not only as follow up in a face-to-face survey.

A response-enhancing measure that might not work so well among ethnic minorities is the use of a longer fieldwork period, since ethnic minorities often have a higher mobility than the general population. As a result, longer or extended fieldwork periods conducted with constant, but not necessarily high intensity throughout the entire fieldwork period often lead to increased nonresponse among these subgroups.

The features mentioned in Table 2.1 focus on increasing contact and reducing refusals in surveys. Reducing non-contact and refusals may be important in obtaining a higher response rate among ethnic minorities, but ethnic minorities do not necessarily differ that much from the general population on these dimensions. For example, simply increasing the number of call attempts can even increase contact rates and subsequently response rates among ethnic minorities to a greater degree than the native population (Feskens et al. 2006; Schmeets 2005). However, the effect of extended call attempts on the response rate among ethnic minorities can also reach its optimum sooner than the native population (Laganá et al. 2013). A big difference compared to the general population is the often higher level of nonresponse due to not being able to participate. This is frequently caused by language problems, functional illiteracy and/or cultural differences. Reducing these causes for nonresponse among ethnic minorities requires the development of tailor-made approaches that take into account the specific problems of the minority population in point. Nonresponse resulting from language problems can be reduced, for instance, by offering the possibility of conducting the interview in the target’s native language. This can be done by providing translated questionnaires, by using bilingual interviewers or family member interpreters.

Nonresponse resulting from cultural differences might be reduced by sending in interviewers with the same ethnic background, who are familiar with the cultural etiquette and customs, or by gender-matching interviewers. The specific cultural knowledge of the interviewers can benefit the survey agency also by highlighting fieldwork periods that could be less than optimal (therefore not planning surveys, for example, at times when the ethnic minority group celebrates important religious festivities). However, one has to be aware of the possible effect this or the previously mentioned response-enhancing measures can have on the measurement error. This will be discussed in more detail in section 2.4.2.

2.4.1 How to evaluate the success of efforts aimed to reduce the inability to participate among ethnic minorities in large scale surveys?

It is not easy to evaluate the success of response-enhancing measures in large-scale surveys. In general, large-scale surveys are not designed as an experiment, which makes it more difficult to assess effects. For instance, some response-enhancing measures, like incentives for reluctant sample units, are only applied to reluctant persons. Consequently, the success of a certain measure is relative and conditional on the survey design and the sample units it was applied to.

The purpose of response-enhancing measures is to increase the response with the aim of hoping to reduce potential bias on estimators as a result of nonresponse. The response rate is, of course, the complement of the level of nonresponse. It is therefore quite common to judge the quality of a survey by looking at the response rate (Biemer & Lyberg 2003). However, nonresponse in itself shouldn't be a problem. As long as nonrespondents and respondents do not differ on the topic being examined, the estimator might just be less precise. Simply increasing the size of the sample could be a solution in this case. The problem arises when respondents and nonrespondents differ systematically on the topic that is being researched, this being referred to as nonresponse bias. The response rate is, however, a poor indicator for nonresponse bias (Curtain et al. 2000; Keeter et al. 2000; Groves and Peytcheva 2008).

As a consequence, to estimate the success of response-enhancing measures aimed to reduce the nonresponse due to inability one should not only check to what degree they have contributed to a higher response rate. The focus should be on the degree to which the measure led to reducing nonresponse bias on an estimator. The problem with this is that the answers of the nonrespondents are, by definition, unknown, and therefore also the degree of nonresponse bias is unknown. Moreover, nonresponse bias happens at the question level, which means that potential nonresponse bias can be different between questions. Sometimes it is possible to determine the size of the nonresponse bias on an estimator via external sources, but this is rather exceptional. Additionally, one should bear in mind that the external sources could themselves be affected by a certain degree of bias. For example, sample estimates based on reference surveys considered to be of higher quality. Besides, this information will often be available only for one or a few of the questions covered in the survey.

To gain a better understanding of nonresponse error and especially of the possibility for nonresponse bias on estimators, it is important to consider aspects such as the differences between respondents and nonrespondents on characteristics that are observed for the entire sample (i.e., paradata), and the relationship between these fully observed covariates and the survey outcome of interest (Andridge and Little 2011). Paradata refers to both process data, such as number or timing of call attempts or interviewer observations, and auxiliary information, like sampling frame data (see Couper 2005; Kreuter 2013; Maitland et al. 2009).

In the last few years, several quality indicators have been developed that take into account not only the response rate, but also one or both aspects mentioned for the assessment of potential nonresponse bias on an estimator (See for instance, Andridge and Little 2011; Särndal 2011; Särndal and Lundström 2010; Schouten et al. 2009; Wagner 2010). However, these quality indicators differ as to how the estimate can be used to correct for nonresponse bias. Furthermore, they differ as to the assumptions related to the missingness pattern (MCAR, MAR, MNAR, see Little and Rubin 2002). They also differ in their ease of use, their purpose of evaluation and the type of paradata required.

An alternative and possibly more insightful manner of evaluating the success of a response-enhancing measures aimed to reduce *inability*-nonresponse is by checking whether the use of the measure resulted in a lower estimated nonresponse bias on a survey outcome according to these quality indicators. In case of an *inability* response-enhancing measure (or more general, a response-enhancing measure targeting a specific subset of sampled persons such as refusers or noncontacts), the effect of the measure can be evaluated by comparing the possibility or the size of estimated nonresponse bias on a survey outcome in the final sample with the sample in which a correction has been applied for the use of the response-enhancing measure. This is basically a combination of two of the five methods – *using rich sampling frame data or supplemental matched data* and *studying variation within the existing survey: nonresponse follow-up studies-* for assessing nonresponse bias described by Groves (2006). However, this requires the availability of relevant paradata about the measure that needs evaluation. When it is not clear, for instance, which respondents joined the sample in the reissue phase, it is, obviously, impossible to determine to what degree the reissue contributed to a reduction of estimated nonresponse bias on a survey outcome.

The studies described in Chapters three and four adopt this approach to evaluate the quality of data collected among four non-Western ethnic minority populations. Several alternative quality indicators are used in order to determine the effect of survey design choices – that is, use of bilingual interviewers and use of a reissue – and that of different survey designs – CAPI or sequential mixed-mode (web-CATI-CAPI) – on the sociodemographic composition of the respondents and on the potential for biased estimates due to nonresponse in surveys among ethnic minorities. In Chapter three, the following quality indicators are used next to the response rate: the representativity indicator, the maximal absolute *standardized* bias (Schouten et al. 2009), final fieldwork disposition codes (De Heer 1999) and partial-R-indicators (Schouten et al. 2010; Shlomo et al. 2009). In chapter four, the following alternative quality indicators are used next to the response rate: the R-indicator, the maximal absolute standardized bias, the fraction of missing information (Wagner 2008; 2010) and partial-R-indicators.

The representativity-indicator or R-indicator is a measure that describes how well the respondent sample reflects (i.e., how representative it is for) the population of interest, based on a certain number of background variables (Bethlehem et al. 2011; Schouten et al. 2009). Obviously, this representativity only applies to the variables included in the model for estimating this measure. 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, p. 154).

The maximal absolute *standardized* bias is an estimate of the upper non-response bias for a hypothetical survey item under the scenario that nonresponse correlates maximally with the selected auxiliary variables (Bethlehem et al. 2011 p. 186).
The final disposition codes refer to the final recorded outcomes to the survey request. They can be used to study the different reasons for nonresponse (see for instance, De Heer 1999). The partial R-indicator is a measure designed to check for the over or underrepresentation of subpopulations in the respondent sample (Schouten et al. 2010). Fraction of missing information 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 (Little and Rubin 2002). The fraction of missing information is defined as the ratio of the between imputation variability to the total variance of the survey estimates. The fraction of missing information can also be used to assess the quality of a sample with respect to potential nonresponse bias for a single item (Wagner 2008; 2010).
In the (quasi)-experimental studies described in Chapter three and four, sampling frame data will be used. However, this is not always available. It is also possible to assess the quality of survey outcomes as far as nonresponse bias is concerned, without using sampling frame data. For example, to obtain insight in the effect of a certain response-enhancing measure, one can extend on the approach introduced by Wagner (2008; 2010). As just mentioned, he proposed to use the fraction of missing information as a method 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. If one were to estimate the average fraction of missing information and the standard deviation based on a large number of target variables and subsequently compare this estimate between the samples that include and exclude the respondents that participated due to the measure assessed one would gain more insight in the effect of the response-enhancing measure. This approach is also investigated in Chapter four.
In a face-to-face survey, one could also collect proxies of sociodemographic characteristics about each sample unit. In this case, at the first contact attempt the interviewer could fill in a short questionnaire in which he/she assesses the type of people living there based on the neighbourhood, the street, the building and possibly the garden. An estimate could be made of, for instance, the socioeconomic category to which the sample unit belongs, the family composition, the type of dwelling, etc. This should be filled in before the interviewer rings the door. This is a way of determining the accuracy of the interviewers' assessments, as eventually both will be available for the respondents: the prior assessment and the answer of the respondent. In this way it may be possible to determine to some degree how the nonrespondents are different from the respondents without having access to sampling frame data. It goes without saying that interviewer observations are more useful in reducing nonresponse bias when information is collected about those characteristics of the sample persons which have a strong correlation with the subject of the study (Kreuter et al. 2010). This, however, mostly implies in-person contact and the method of using interviewer observed paradata certainly has its challenges and limitations with respect to nonresponse adjustment (see for instance, Sinibaldi et al. 2013; Olson 2013 or Kreuter and Olson 2013).

2.4.2 Response-enhancing measures: the trade-off between representation and measurement

Some response-enhancing measures are being taken in order to reduce nonresponse among ethnic minority members that otherwise are not able to take part in the survey. The expectation is that the persons that are not able to participate without these measures might differ with respect to the topic of interest. In the case of survey research among minorities, response-enhancing measures target mostly the reduction of nonresponse bias due to linguistic problems, illiteracy and/or cultural differences. After all, the fact that a person cannot read, doesn't speak the (majority) language or is culturally socialized in a different way can have a serious influence on their participation in society and their perspective on it. It is important to realise here that *not being able* to participate can also include ethnic minority members that *will not* participate unless their cultural etiquette is taken into account with respect to the survey request.

The downside of these measures meant to reduce `not able` nonresponse is the chance for increased measurement variability. For instance, if one or more translations of the survey are necessary, there is a chance of translation errors or translation induced differences (Harkness and Schoua-Glusberg 1998; Harkness et al. 2004). In turn, the use of (*bilingual*) interviewers with the same ethnic background as the respondent can also lead to increased measurement variability (Anderson et al. 1988; Davis 1997; Finkel et al. 1991; Van Heelsum 2013; Van't Land 2000).

An important question in this sense is then at what point the impact of the response-enhancing measures on the nonresponse error ceases to outweigh the measurement error introduced. This is a tough question to answer. However, there are a number of considerations that can facilitate the choice whether to implement such a measure in research among ethnic minorities. A first important consideration concerns the surveyed population. Does the survey mean to obtain a representative image of the general population in which ethnic minorities are represented, or is it meant to get a representative image of the ethnic minorities? In the first case, the effect on nonresponse error will be relatively small for a rather high investment, while taking into account an increased probability of measurement variability. Although even in those circumstances other aims -such as the aim to report important indicators about different subgroups- may increase the necessity of employing a response-enhancing measure (see for example Stoop 2014).

A second consideration would be whether the survey topics consist of more structural, factual issues or of `softer` issues such as questions about attitudes. With more structural questions, the impact of response-enhancing measures such as the introduction of interviewers with a shared ethnic background on the respondents' answers appears to be smaller than with `softer` issues (see Chapter 5). Measurement differences due to translation can, however, surface in all situations. Furthermore, one should also realise that response-enhancing measures increase the costs of a survey

It is important to be aware in the planning stage of the survey of the ways in which a response-enhancing measure can influence the answers of respondents. Is it possible to take precautions in order to minimise or correct this effect, like, for instance, a post-hoc correction of the data for possible measurement effects introduced due to the response-enhancing measure? However, this needs a clear hypothesis and a design that makes the correction possible.

Several types of procedures can be used in order to reduce the possible influence of measurement differences and increase comparability introduced by response-enhancing measures. Worth mentioning are cognitive interviews (see, for instance Beatty and Willis 2007), strict translation protocols (Harkness 2007), careful and user-friendly questionnaire design (Dillman 2007) and procedures focussing on the characteristics of questions (e.g. Saris and Gallhofer 2014).

2.5 Comparability, timeliness and costs concerns in surveys among ethnic minorities

In this section we focus on two important Eurostat criteria for quality – comparability and timeliness- in assessing survey data collected among ethnic minorities (see Eurostat 2000). Comparability refers to the degree to which one can compare data originating in different surveys, periods or countries or data concerning different target groups. Timeliness means the speed with which data becomes available. However, the quality of survey data is not a concept that should be assessed independently. Other factors, such as the burden, cost, professionalism or design constraints determine to a large extent how one ought to assess the quality of survey data. Therefore we shall also take a closer look at cost and cost-related conditions like design constraints at the end of the current chapter.

2.5.1 Threats to comparability: using the cross-cultural perspective

Cross-cultural and cross-national comparative research offer an interesting perspective on the quality criterion of comparability which can be applied for the assessment of the quality of survey data among ethnic minorities. This type of research specifically studies the multitude of errors and biases that can complicate or even invalidate the comparability of cross-cultural or cross-national collected data (see, for example Davidov et al. 2011; Harkness et al. 2010; Stoop et al. 2010).

In cross-cultural or cross-national research, three types of measurement bias that can threaten the validity of comparisons are commonly distinguished. These are construct bias, item bias and method bias (Van de Vijver 2011). Construct bias occurs when the requirement of construct equivalence is not met. This can happen when non-identical constructs are measured across cultures or countries, or when there is only a partial overlap of the construct between the cultures or countries. Construct bias is introduced at the level of the measurement instrument designed to capture the theoretical concept. Item bias happens at individual question level and occurs when translations of questions lead to differences in question meaning or ambiguity for different groups. Item bias can also be the result of cultural specifics which can be viewed as a form of differential item functioning or DIF (Mellenbergh 1989). DIF is a term that stems from education testing and happens when persons of equal capability or intelligence arrive at different capability or intelligence scores based on the specific wording of an item.

Method bias happens at survey level and can be introduced by a variety of factors which are distinguished in the following three categories: *incomparability of samples*, *administration bias*, and *instrument bias*. *Incomparability of samples* refers to differences in the sample composition with respect to important sociodemographic characteristics of the respondents. *Administration bias* refers to bias that is introduced as a result of differences in how the questionnaire is administered (e.g., interviewer effects, presence of others during the interview, interviewer characteristics, differences in interview language), differences in questionnaire design, differences in mode of administration, and so on. *Instrument bias* refers to bias that is introduced as a result of differences in familiarity with being interviewed, but also differences in cultural specific answer strategies.

Besides measurement biases that can threaten the validity of cross-cultural comparisons, the existence of different survey realities between ethnic minority groups that are more or less out of the survey designers' control, such as coverage issues, differences in nonresponse bias or the lack of suitable sampling frames, can also seriously threaten the validity of cross-cultural comparisons.

2.5.2 Lessons from cross-cultural research: taking comparability into account

In much cross-cultural research, the purpose is to compare different groups with one another on different indicators, like sociocultural integration or socioeconomic position. The considerations above show that the comparability of data can be disrupted in many ways.

Several types of procedures that can be used in order to detect and to reduce the possible influence of construct and it bias on data comparability have already been mentioned in section 2.3.4. However, detecting and reducing the potential influence of method bias on data comparability is more difficult, because method bias is a factor at the survey level, while the comparability of data concerns the question level. It is possible, for instance, that method bias should impact one question, but not another. The degree to which, for example, the presence of an interviewer generates socially desirable answers may differ between questions and, possibly, between respondents. How method bias can invalidate cross-cultural comparisons is investigated in Chapter six.

On the other hand, introducing method bias in order to adapt to the reality of surveying ethnic minorities, for instance by using bilingual interviewers with the same ethnic background and interviewers of a different origin can be necessary in order to increase representativity. At the same time, it is hard to predict in advance the influence on data comparability of factors that are difficult or impossible to control, such as differences in the coverage capacity of a certain sampling frame for one or several ethnic minority groups.

Lynn (2003) recognizes the same problems in conducting cross-national research, specifically how data comparability can be put at risk by differences between countries insofar as the survey reality and the ways of conducting the surveys are concerned. The assumption is that large differences in the conditions, design and method of conducting surveys between countries affect the comparability of survey results. Lynn (2003) uses this insight for developing a series of quality standards for cross-national survey research. Explicit documentation of the choices and differences in the conditions, design and method of conducting surveys is an essential step in this process.

Documenting the survey or social reality of different ethnic minority groups also helps in the assessment of survey data quality among these groups. This way it becomes explicit in what way the survey reality of ethnic minorities and the survey design choices can threaten the comparability of data. An important difference with Lynn (2003) is that cross-cultural surveys, unlike cross-national surveys, are often set up in a single country using a single fieldwork agency. This means that threats to the comparability of data originating from various fieldwork agencies with different survey cultures, interviewer training programmes, payment schemes or experiences should not be a factor to take into account. Even though, in some situations parts of the fieldwork could be subcontracted to another fieldwork agency.

Obviously, it is not always possible to assess the effects that differences in survey design choices and survey reality may have on comparability. However, documentation of the differences in survey reality and research objective will contribute to determine the survey design choices and will offer insights for interpreting and comparing the results. Furthermore, it is very valuable for other researchers and (re)users of the data if one describes the choices made regarding comparability or, even better, if one includes variables that help model the effect in case the user needs the data for a different purpose.

2.5.3 Timeliness and costs considerations

The timeliness with which data becomes available is often very important. Society is changing constantly, which means that data which is available rather late can lose its relevance. However, when research focuses on delivering data very quickly, accuracy can be put at risk. An example could be the choice of a very short fieldwork period, without any follow-up of nonrespondents, or else the use of a limited number of contact attempts, insufficiently spread out over the fieldwork period.

Consequently, it is important to determine to what extent the target population imposes restrictions on how quickly data can be collected. In this sense, using bilingual face-to-face interviewers is inevitable when conducting survey research among ethnic minorities with a high frequency of functional illiteracy and/or linguistic problems. However, this method of data collection is relatively more time-consuming given the fact that any interviewer is only capable of a limited amount of contact attempts in a fixed period. Other methods of data collection are not or less conditioned by the capacity limitations of interviewers. For instance, all selected sample persons can be invited at once to take part in a web interview. Yet the known problems of a certain target population will determine how adequate such methods of data collection are. In this context, it is also important to differentiate between surveys among the general population of which ethnic minorities are a part and surveys exclusively among ethnic minorities. In the first case, the choice of a data collection method without or with lower capacity restrictions may have a smaller impact on the accuracy of data in correlation with the increase in timeliness than in the second case.

Timeliness of delivery can also be at stake when combinations of data collection methods are used. In Chapter 4 we investigate this issue to a greater extent when we compare a sequential mixed-mode survey – Web-CATI-CAPI – with a single-mode CAPI survey that were both conducted among non-Western populations in the Netherlands as part of a large survey design experiment.

Furthermore, there are certain cost-related considerations that are frequently overlooked in survey research among ethnic minorities. One needs to consider, for instance, the extra expenses related to the use of tailor-made response-enhancing measures. Translation of questionnaires and particularly the use of bilingual interviewers with a common ethnic background make for a time-consuming and costly process. More often than not, these interviewers have to be recruited and trained specifically for a survey and, in case of inadequate results, high turnover must be avoided by using both financial and softer motivation-enhancing incentives.

In this context, applying a combination of cheaper modes can lead to a distorted image of the actual costs of a survey. Face-to-face interviews are the most expensive form of data collection, therefore reducing or avoiding this method of data collection might, at first sight, generate a serious reduction in cost. Especially in the case of large surveys, this measure can lead to economies of scale. However, an aspect often overlooked in such cost-efficiency calculations is the fact that web interviews and mail questionnaires have to be shorter than face-to-face interviews. This is for good reason, as longer web interviews or mail questionnaires lead to lower response rates and a higher loss of data quality, for instance by satisficing (Galesic and Bosnjak 2009). The downside of short questionnaires is, however, that less information is obtained from the respondent. From this perspective, the savings obtained by using cheaper data collection modes might actually be substantially smaller than expected. These considerations of time and cost are particularly relevant when sample sizes are relatively small and the known survey difficulties in connection to specific target populations require the use of a face-to-face mode.

This is will be well illustrated by the results of an experiment conducted among ethnic minorities in which the quality and costs of a single-mode CAPI survey was compared with the quality and costs of sequential mixed-mode surveys (see Chapter 4).

This chapter offers an informative overview of the pitfalls and challenges of collecting survey data among ethnic minorities by placing these in the context of the TSE paradigm. An especially common pitfall is the mismatch between the social reality of ethnic minorities and the standard set of survey practices and survey design features. When they do not correspond well, they affect coverage error, nonresponse error, measurement error and post hoc adjustment in surveys among ethnic minorities to a large extent. Furthermore, the dilemma between nonresponse and measurement error needs careful consideration. Not only when designing a survey among ethnic minorities, but also in assessing the accuracy of the estimates when tailor-made response-enhancing measures were used. The impact of survey design and response-enhancing measures on the representation of Non-Western minority populations in the Netherlands are the topic of study in Chapters three and four. The impact of survey design and response-enhancing measures on the measurement of substantive variables among Non-Western minority populations in the Netherlands will be the topic of study in Chapter five. At the same time, this chapter provided insight in other important quality issues pertaining to survey data collected among ethnic minorities. And, although comparability falls outside the TSE framework we hope that we have demonstrated why it is an important quality indicator to take into consideration in surveys among ethnic minorities. The cross-cultural comparability of survey outcomes among non-Western minorities in the Netherlands will be the topic of study in Chapter six.

3 The effect of different survey designs on nonresponse in surveys among non-Western minorities in the Netherlands

The present chapter investigates the impact of survey design choices on the representativity and the potential for nonresponse bias on survey estimates of eight sub-surveys conducted among non-Western minorities in the Netherlands. This study utilizes fieldwork disposition codes in conjunction with the R-indicator and maximal absolute standardized bias to show the impact of survey design choices – such as the period and length of fieldwork, the use of bilingual interviewers, the number of face-to-face call attempts and a reissue of nonresponding sampled persons – on the potential for nonresponse bias on survey estimates. Partial R-indicators are used to detect which sociodemographic subgroups contribute the most to a nonrepresentative response, conditional on ethnic group and survey design. The results indicate that long fieldwork periods increase the potential for nonresponse bias on survey estimates among non-Western minorities due to moving and that the timing of fieldwork has an impact on the number of sampled persons who are unavailable during the fieldwork period. Furthermore, the use of bilingual interviewers is necessary to conduct a survey among Turkish and Moroccans due to language problems; otherwise the potential for nonresponse bias on survey estimates can be quite severe. Also, the use of a reissue phase reduces the potential for nonresponse bias on survey estimates in surveys among non-Western minorities in the Netherlands. Finally, partial R-indicator analyses provide further insight on how future surveys can be improved in order to reduce the potential for nonresponse bias on survey estimates among each of the four non-Western minority groups[1].

3.1 Introduction

In general population surveys, non-Western minorities – or ethnic minorities as they are sometimes referred to – tend to be underrepresented (Feskens 2009; Groves and Couper 1998; Schmeets 2005; Stoop 2005). At the same time, there is a great need for specific information about this group, especially on issues such as socioeconomic and sociocultural integration in the Netherlands and elsewhere (Bijl and Verweij 2012). That is why separate surveys among non-Western minorities continue to be necessary. 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 non-Western minorities. It is therefore of great importance to determine which strategies are effective for surveying non-Western minorities, while maintaining a certain level of quality and minimizing the costs.

[1] This chapter has been published as Kappelhof, J.W.S. (2014). The effect of different survey designs on non-response in surveys among non-Western minorities in the Netherlands. Survey Research Methods, 8, 2, 81-98

This chapter sets out to investigate how different survey design choices affect the composition of the *response* sample (i.e., the composition of the group of respondents) and how this might relate to the occurrence of nonresponse bias on survey estimates in surveys conducted among non-Western minorities in the Netherlands. We shall compare eight sub-surveys – four separate sub-surveys in two different survey rounds – that vary in these choices and we shall try to ascertain which set of design choices leads to the sample with the lowest potential for nonresponse bias on survey estimates.

A standard measure for judging the quality of a response sample is still the response rate, despite the fact that it is not a direct measure of nonresponse bias (Biemer and Lyberg 2003, Groves 2006). In the last few years several other quality indicators have been developed that – under assumptions – provide a more direct insight in the existence of nonresponse bias and allow us to estimate its size (see for instance Andridge & Little 2011; Särndal 2011; Särndal and Lundström 2010; Schouten et al. 2009; Wagner 2010). In this chapter, next to the response rate, we shall make use of two methods to evaluate the quality of the response samples of both surveys among non-Western minorities and its potential for nonresponse bias on survey estimates.

The first method is based on studying different reasons for nonresponse by analysing the final disposition code of the sample units (see for instance, De Heer 1999). The second method utilizes the representativity indicator (R-indicator) and the related maximal absolute standardized bias estimator ($\widehat{Bm}$) to study nonresponse (Bethlehem et al. 2011; Schouten et al. 2009). We also analyse the impact of separate survey design choices, such as the number of face-to-face contact attempts, the reissue of non-responding sampled persons and the use of bilingual interviewers with a common ethnic background. To this end, we use the R-indicator and $\widehat{Bm}$ to show the impact of these design choices on the quality of the response sample. We conduct a detailed analysis of the under- and over-represented sociodemographic subgroups within each survey design, separately for each minority group, using partial R-indicators (Schouten at al. 2013). This will allow us to further develop tailor-made approach strategies for future surveys among non-Western minorities in the Netherlands.

The chapter starts with a brief overview of the main data collection difficulties resulting in nonresponse when surveying non-Western minorities. The data and methods section describes the surveys, the survey design choices and the methods used to answer our research aim. This is followed by the results of the analysis and the subsequent conclusion and discussion.

3.2 Why are non-Western minorities underrepresented in population surveys in the Netherlands?

In 2011, non-Western minorities made up about 11% of the population in the Netherlands (CBS-statline). Statistics Netherlands uses the following official definition: “Every person residing in the Netherlands of whom one or both parents were born in Africa, Latin-America and Asia (excluding Indonesia and Japan) or Turkey (Reep 2003)[2].

The main reason for the underrepresentation of non-Western minorities in population surveys in the Netherlands is nonresponse. One can make a distinction between direct causes and correlates for nonresponse on the one hand, and between characteristics of the person and the survey design features on the other hand. 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 tend to live more often 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 the direct causes of nonresponse are addressed may reduce the specific nonresponse among non-Western minorities. Language problems stop being a problem when one of the design features is a translated questionnaire. Functional illiteracy ceases to be a problem when the interviews are conducted by interviewers who read out the questionnaire. Also, the use of the telephone for interviews increases the number of refusals among non-Western minorities to an incomparable degree in comparison to native Dutch or to a face-to-face mode and should therefore be avoided (Schothorst 2002; Korte and Dagevos 2011).

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: they do not only speak the language, but are also aware of the proper etiquette to approach sampled persons. An often overlooked cause 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 summer. Sometimes, there is also a mismatch between religious holidays of ethnic groups and the way the fieldwork agency plan their fieldwork (Kemper 1998; Schothorst 2002; Veenman 2002).

Sampling frame errors and especially undercoverage provide other reasons why non-Western minorities are underrepresented in population surveys in the Netherlands. Undercoverage is what happens 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 munici-

[2] 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 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. It mainly involves persons born in former Dutch-Indie (Indonesia) and employees working for Japanese companies with their families.

pal personal data records database (population register) as a sampling frame. Both frames suffer from frame errors, such as moving of the sample unit, no known address of the sample unit, slow registration of the sample unit or death of the sample unit. Some of these causes seem to occur far more often among non-Western minorities, such as moving or no known address of sample units (Feskens 2009; Kappelhof 2010).

3.3 Data and Methods

3.3.1 Data

The 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 which started in 2006 and was repeated in 2011. In the present study both face-to-face CAPI survey rounds are included (SIM2006 and SIM2011).

In both SIM-rounds Statistics Netherlands drew a random sample of named individuals from each of five mutually exclusive population strata; Dutch of Turkish, Moroccan, Surinamese, and Antillean[3] descent and the remainder of the population (mostly native Dutch) living in the Netherlands, aged 15 years and above. The present study focusses on how different survey design choices affect the potential for nonresponse bias on survey estimates in surveys conducted among non-Western minorities. This is why the samples containing native Dutch are excluded from this study resulting in eight samples for analysis.

The official definition of Dutch of Turkish, Moroccan, Surinamese, and Antillean descent includes persons that were either born in Turkey, Morocco, Surinam or the Dutch Antilles[4] or have at least one parent who was born there. In case 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.

Both SIM-rounds used the population register as a sampling frame and the same stratified two stage probability sampling design in all four population strata. In the first stage municipalities were selected and in the second stage 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 (self-selecting); 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 3.1).

[3] Including Aruba.

[4] or Aruba.

Table 3.1
Gross sample sizes per ethnic group and survey year across municipality strata

	Turkish		Moroccans		Surinamese		Antilleans	
	2006	2011	2006	2011	2006	2011	2006	2011
Large municipalities	802	554	1,218	812	1,563	1,020	867	695
Midsize municipalities	928	727	771	674	714	662	947	945
Small municipalities	432	284	401	254	401	248	398	334
Total	2,162	1,565	2,390	1,740	2,678	1,930	2,212	1,974

In this study we used the fieldwork and response data files from SIM2006 and SIM2011. The fieldwork data files 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, sex, first or second generation immigrants, municipality, etc. Process data and auxiliary information, also known as paradata, are potentially useful for increasing participation, for nonresponse adjustment or for evaluating potential nonresponse bias (Couper 2005; Kreuter 2013; Maitland et al. 2009). The response data files contain the answers of the respondents to the survey questions, but also interviewer observations about respondents, such as their ability to speak Dutch.

Survey design choices and response-enhancing measures

There are differences in the survey design between both SIM-rounds with respect to the fieldwork and the questionnaire. In SIM2006, the main part of the fieldwork lay outside the winter period, whereas for the SIM2011 survey the main part of the fieldwork was conducted during the winter period. The length of the SIM2006 fieldwork was also about twice that of the SIM2011 measurement: nine months versus five months. Also, the fieldwork agencies differed across rounds. Bureau Veldkamp conducted the fieldwork in 2006 and Gfk Netherlands[5] in 2011.
The main difference about the questionnaire between the surveys resided in the length. The research topics were identical, but the questionnaire length was reduced. The reason for this reduction was based on interviewer reports after the completion of the SIM2006 survey, but also on opinions of fieldwork experts and experts on minority research (Feskens et al. 2010). They all believed the questionnaire was too long which could potentially harm the response rate. This resulted in a reduced questionnaire length between the SIM2006 and the SIM2011 measurement from an estimated, based on CAPI timers, average of 55 minutes to 44 minutes.

[5] Gfk also made use of a subcontractor (Labyrinth) to ensure enough interviewers with a shared ethnic background were available to conduct the fieldwork among all ethnic groups.

Response-enhancing measures such as the use of incentives and advance letters have a proven positive effect on the response rates (Dillman 2007; Groves and Couper 1998; Singer et al. 1999; Singer et al. 2000; Singer 2001). These measures may therefore also affect the response composition and the quality of the response sample.

The type of measures that were used varied between the two SIM-rounds and ethnic groups. There were also differences in the extent to which the same measures were used in 2006 and 2011, and in the ethnic groups. An unconditional nonmonetary incentive (stamps) was used in SIM2006 among all groups, whereas no unconditional incentives were used in the SIM2011 measurement.

Conditional incentives were used among all groups in both surveys. All respondents received a gift certificate (€10) after completion. In SIM2011, respondents were also offered the option to donate €10 to charity.

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 mostly found among the Turkish and Moroccan population (Kappelhof 2010). For both SIM2006 and SIM2011, auxiliary information about ethnicity, age, sex, 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 of the sampled persons. Two types of tailoring were used to increase response. They mainly have to do with anticipated language problems, but also with anticipated cultural differences. Research has shown that a greater cultural familiarity due to the common ethnic background between 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 SIM2006 and SIM2011, but only among Moroccan and Turkish. Also 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 recorded 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 common ethnic background. Both surveys used interviewers with a shared ethnic background with the sampled person, but the intensity in which they were used varied between SIM2006 and SIM2011 *and* between target groups.

In both SIM-rounds *bilingual* interviewers with a common ethnic background approached sampled persons of Moroccans or Turkish origin. In SIM2006 there was a limited and systematic use of bilingual interviewers with a common ethnic background among a part of this group. They mainly contacted older, first-generation immigrants who lived in the larger cities, because that is where the language problems were mostly anticipated. For respondents that were interviewed by *non-bilingual* interviewers without a common ethnic background, the translated questionnaire was also made available.

The questionnaire could be shown on request of the respondent or in case a question posed in Dutch was unclear to the respondents. Interviewers with a common ethnic background were hardly used at all among sampled persons of Surinamese or Antillean origin in the SIM2006 study.

In SIM2011, all sampled persons of Moroccan or Turkish origin were contacted by a bilingual interviewer with a common ethnic background. In SIM2011 about half of the sampled persons of Surinamese or Antillean origin were approached by interviewers with a common ethnic background, the other half were approached by either Dutch interviewers or interviewers with another ethnic background. The allocation of Surinamese and Antillean sample units to interviewers with a common ethnic background was based on the availability of an interviewer with a common ethnic background in the area.

In 2006 and 2011, potential respondents could call a toll free number in case of questions or to reschedule an appointment for an interview. Finally, interviewer bonuses to increase interviewer productivity were used in SIM2006, but not in SIM2011. Unfortunately, there was no information available on the identity of the interviewers who received these bonuses in 2006, so as to analyse the effectiveness of this measure.

The reality of fieldwork: deviations from the planned survey design.

Both SIMs used a responsive design approach where non-responding sampled persons in the first phase of fieldwork are taken 'out of the field' and reissued again by the fieldwork agency (Groves and Heeringa 2006). This approach provides the opportunity to introduce other design choices after the first phase, such as an increased incentive or another interviewer.

In the first phase of SIM2006, a minimum of four contact attempts (CA) had to be made to a sampled person before the sampled person could be registered as a noncontact and returned to the fieldwork office for potential re-issuing. The CAs had to be made on different days and at different times in the day.

In SIM2011, there had to be at least three CAs on different days of the week and at different times during the day before the sampled person could be registered as a noncontact and returned to the fieldwork office. However, interviewers were encouraged to conduct more CAs. Only after three unsuccessful CAs in the first phase, the interviewer was allowed to try and reach the sampled person by telephone (if available) and set up an appointment or leave a "sorry I missed you" card.

The way unsuccessful sampling units were selected to be reissued in the second phase varied between both SIMs. In SIM2006, the planned second phase of fieldwork involved only the reissue of soft refusals and noncontacts among underrepresented non-Western minority subgroups, such as young males living in urban areas. These reissued sampled persons were offered the same conditional incentive worth €10 and a minimum of four CAs had to be made by another interviewer.

Unfortunately, during the second phase of the SIM2006 fieldwork not all sample units selected for re-issuing were re-contacted with a minimum of four contact attempts for noncontacts. The difference in selection and reissue of unsuccessful sample units back into the field was based on the availability of another interviewer in the area and costs. This meant that, if a sample unit was selected to be reissued but no other interviewer was available in the area, none would be sent in case there were less than three reissued sample units.

In total 1,143 sample units were selected for reissuing in 2006. Unfortunately fieldwork ended before all sample units selected to be reissued were actually reissued or re-contacted at least four times. This resulted in 522 second phase sampled persons that were either not reissued or where no final disposition code was achieved. Only for 621 sample units a final disposition code was declared (see Table 3.2).

In SIM2011 the plan was to select all first phase nonrespondents and to reissue them for the second phase. A minimum of three face-to-face contact attempts had to be made by another interviewer. Furthermore the amount of the promised or conditional nonmonetary incentive (gift certificates) was increased from €10 to €15.

Unfortunately, again, due to time constraints, only very few sample units were actually re-contacted by another interviewer (Table 3.2). In this case, the difference in selection and reissue of unsuccessful sample units was based on the availability of another interviewer in the area within the remainder of the fieldwork period. In case no other interviewer was available in the area, the original interviewer had to conduct at least six contact attempts.

Table 3.2

Sample units selected for face-to-face CAPI reissue in SIM2006 and SIM2011

	Number of nonresponding sampled persons selected for reissue		Number of sampled persons not reissued or with no declared final disposition code		Number of sampled persons reissued with a final disposition code	
	2006	2011	2006	2011	2006	2011
Turkish	250	346	108	288	142	58
Moroccans	217	242	102	234	115	8
Surinamese	413	453	214	227	199	226
Antilleans	263	485	98	303	165	182
Total	1,143	1,526	522	1,052	621	474

3.3.2 Methods

The analysis of data from nationwide, cross sectional surveys among hard to reach populations for which specific measures were undertaken imposes limits on the use of analysis methods, such as logistic regression. Both SIMs are not set up as an experiment to assess the effectiveness of separate response-enhancing measures on the probability of participation among various sociodemographic subgroups. They were designed to be as efficient as possible in increasing the probability of response among various, very difficult to survey populations by using auxiliary information available on the sampled persons. This meant, for instance, a non-random allocation of sampled persons with specific characteristics to ethnic interviewers in SIM2006. Also, in both SIM-rounds only certain nonresponding sampled persons were selected and actually reissued. As a result, the effect of sociodemographic variables such as age, immigration generation, municipality size, ethnic group on the odds to participate is confounded with the non-random allocation of a (bilingual) interviewer with a common ethnic background and with whether or not a sampled person has been reissued, in which case they were usually contacted by a more successful interviewer.

Another potential confounding factor is the possible change in perception of surveys and in general willingness to participate in surveys that may take place in the interval between both survey rounds among the hard to reach minorities. For instance, in the five year interval, a continuing shift towards the right was noticeable in Dutch society, combined with the rise of a more populist discourse on migrants in the Netherlands. This might negatively affect the willingness to participate of non-Western minorities.

The representativity-indicator (R-indicator) and the maximal absolute standardized bias are quality indicators that allow for a comparison between surveys using different, targeted designs and/or a comparison across time (Schouten, Cobben and Bethlehem 2009). Recently, both indicators have been developed as a result of a large European project to assess the effects of nonresponse on the quality of statistics (RISQ-project.eu). These indicators are not dependent on a random allocation of sample units, but allow for an assessment of the quality of the response sample in which targeted response-enhancing measures were used. They also allow for an estimation of the impact of separate response-enhancing measures on the quality of the response sample.

The following two approaches, which we will present in more detail, are used to ascertain the quality of the response sample. The first approach is the final disposition code of the sample unit and the second approach is the representativity indicator (R-indicator) in conjunction with the maximal absolute standardized bias ($\widehat{Bm}$). Furthermore, the impact of the different survey designs on the balance of the response across different subgroups in each ethnic group will be assessed via partial R-indicator analysis (Shlomo et al. 2009). These results will be used to gain insight on how to further improve fieldwork. It is important to note that the study of underrepresented subgroups in a response sample, given a certain survey design, is different from estimating the effect of separate response-enhancing measures on the propensity to respond among various subgroups.

Final disposition codes

The complement of the response rate is the nonresponse rate. The nonresponse rate can be used to gauge at the potential for nonresponse bias, specifically the underlying mechanism for nonresponse (Groves 1989; Lynn et al. 2001; Stoop 2005). Refusing to participate or not being able to participate are two different causes of nonresponse and offer an additional insight on the potential for nonresponse bias. Process or paradata information can be used to evaluate how well a specific set of survey design features is able to accurately survey our population of interest.

One way to gain insight is by analysing the final disposition code of nonresponding sample units. There are several main reasons for nonresponse, such as refusal, noncontact, not available, not able, language problems, moved, etc. Each of these reasons may be caused by a specific difficulty of surveying non-Western or ethnic minorities in The Netherlands, which in turn provides insight in the way the response sample reflects our population of interest. Furthermore, this specific information can be used to assess the probability of nonresponse bias for survey items if there is a known relation between the topic of interest and a specific cause for nonresponse. An example might be the correlation that exists between employment status and language problems or functional illiteracy. If persons are not able to speak and/or write Dutch, their chance on having a job in the Netherlands decreases. Another example would be the correlation between home ownership and high mobility. It is fair to say that the probability of a highly mobile person being a home owner is rather low. Nonresponse due to moved sample units varies between non-Western minorities and native Dutch. Non-Western minorities, especially Antilleans, move around more often than native Dutch (Feskens 2009). This difference will increase if the fieldwork period is longer. So, if a specific set of survey design choices leads to an underrepresentation or exclusion of certain subgroups, the response sample will not give an accurate reflection of the population of interest. Survey design choices such as the decision not to use bilingual interviewers or translated questionnaires will cause a high nonresponse rate due to language problems or functional illiteracy. Even if the composition of the response sample is similar to the population of interest with respect to correlated background characteristics, such as age and immigration generation, the underrepresentation of subgroups with language problems may cause biased estimates.

Analysing final disposition codes is straightforward and the appeal of this method is the ease with which it can point out potential nonresponse biases as well as provide insight for the development of new tailor made approach strategies. Furthermore it uses more information than just the response rate in order to judge the quality of the response sample.

Representativity-indicator and the maximal absolute standardized bias

The representativity-indicator (R-indicator) is a measure that describes how well the response sample reflects (i.e., how representative it is for) the population of interest, based on a certain number of background variables (Bethlehem, Cobben and Schouten 2011; 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. 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. Obviously, the individual response propensities are unknown and the fewer distinct strata used to estimate the average response propensities, the less informative the R-indicator tends to be. 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, p. 154). In essence one can view it as a measure that uses the variability between nonresponse adjustment weights. The larger the variability is in nonresponse adjustment weights, the lower the R-indicator will be.

The R-indicator is useful in a variety of ways. First of all, it allows for the comparison of surveys, provided the same variables are available to estimate the model for each survey. Secondly, it is easy to interpret. It is one single estimate between zero and one (or 0% and 100%). Zero means a complete lack of representativity and one means a perfect fit. Thirdly, it can be used to monitor the progress of fieldwork and make more informed decisions on when and how to intervene. Fourthly, it can assist in designing a survey and provide an estimate of the quality while constraining other important parameters such as time and budget. Finally, Schouten et al. (2009, p. 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."

For the estimation of the maximal absolute standardized bias ($\widehat{Bm}$) Schouten et al. (2009) make use of the proof provided by Bethlehem (1988) and Särndal and Lundström (2005) that the bias of the Horvitz-Thompson estimator is approximately equal to the population covariance between survey items and the response probabilities divided by the mean response probability. The following equation [Eq. 1] from Bethlehem (2011) shows the relation between the (estimated) average response probabilities ($\hat{\bar{\rho}}$), the R-indicator $\hat{R}(\hat{\rho})$, the estimated standard deviation of the survey item, $\hat{S}(y)$ and the maximal absolute bias $\widehat{B_m}(\hat{\rho}, y)$.

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

For an unambiguous comparison, Bethlehem et al. (2011) propose to use a hypothetic survey item with a known and equal variance, for example $\widehat{\mathrm{S}}(\mathrm{y})$ =1. This results in the estimated maximal absolute *standardized* bias [Eq. 2]:

$$\widehat{B'_m}\left(\hat{\rho}, y\right) = \frac{\left(1-\hat{R}\left(\hat{\rho}\right)\right)}{2\hat{\bar{\rho}}} \tag{2}$$

The $\widehat{B_m}(\hat{\rho}, y)$ presented in equations (1) and (2) is an estimate of the upper non-response bias for a hypothetical survey item under the scenario that nonresponse correlates maximally with the selected auxiliary variables (Bethlehem, Cobben and Schouten 2011, p. 186).

Unconditional and Conditional partial R-indicators

Sometimes certain sociodemographic subpopulations can be expected to have a different position or opinion on important research topics such as having a job or the attitude towards sociocultural integration. When they are underrepresented in the response sample, the results with respect to these research questions may be biased. It is therefore important to see how such subpopulations are represented in the response sample, given a certain survey design. We shall use partial R-indicators to check for the over- or underrepresentation of subpopulations in the response sample (Schouten et al. 2010; Schouten et al. 2011; Schouten et al. 2013; Shlomo et al. 2009). These subpopulations 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 to the overall lack of representativity of the response sample.

There are unconditional and conditional partial R-indicators for discrete variables. The unconditional partial R-indicator on a variable level can be used to compare between surveys (Shlomo et al. 2009, p. 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 nonnegative and bounded above by 0.5 (Schouten et al. 2011, p. 236).

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, p. 237). It tries to isolate the part of the nonrepresentative response that is attributable to a specific variable. The conditional partial R-indicator on a variable level can take on any value in the interval [0, 0.5.]

Both partial R-indicators can also be calculated on a category level to ascertain the contribution to the lack of representative response separately for each category. The values of the unconditional partial R-indicators on a category level can be positive and negative. A negative value indicates an underrepresented category and a positive value indicates an overrepresented category. The unconditional partial R-indicators on the category level may take values between -0.5 and 0.5, where 0 means no contribution (Schouten et al. 2011, p. 236).

The values of the conditional partial R-indicator on the category level are always positive 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; the values range from 0 to 0.5.

3.4 Results of the different quality indicators

3.4.1 Final disposition codes: response rate and nonresponse composition

In this part, the paradata used are the final disposition code of the sample units.

Table 3.3 presents the breakdown for ethnicity in final disposition code of the sample units for each survey. Here we use the AAPOR definition 1 (RR1), the minimum response rate[6] (AAPOR, 2011). Among Moroccans, there is a significantly higher response rate in SIM2011 compared to SIM2006. The other three ethnic groups show no significant difference in response rates over time. This indicates that the survey design used in the SIM2011 measurement might have successfully counteracted the general trend of decreasing response rates (De Heer and De Leeuw 2001).

When we use the information from the final disposition code to judge which of the samples reflects the population of interest, we can draw four general conclusions with respect to the (planned) different survey design choices. First of all, the survey with the longest fieldwork period (SIM2006) suffers more from an outdated sample frame due

[6] This definition was slightly adapted for the Dutch situation since the AAPOR guidelines do not provide for In Person Surveys of Specifically Named Persons.

to *moving* (Table 3.3). This can cause quite significant nonresponse among non-Western minorities. The second conclusion is that the targeted use of bilingual interviewers with a common ethnic background in SIM2006 still resulted in a higher exclusion of sampled persons among Turkish and Moroccans due to *language problems* compared to the complete use of bilingual interviewers with a common ethnic background in the SIM2011 survey. Thirdly, the timing of the fieldwork in SIM2006 caused a greater number of Turkish and Moroccans sampled persons to be *unavailable during fieldwork*, despite the longer fieldwork time and the larger number of reissued unsuccessful sampled units.

Table 3.3
Final disposition code (in %) per ethnic group per survey year

	Turkish		Moroccans		Surinamese		Antilleans	
	2006	2011	2006	2011	2006	2011	2006	2011
Interview (RR1)	52.9	52.1	43.8*	48.0*	40.1	41.0	46.2	44.2
Moved	5.7	2.9	5.8	4.2	6.6	4.7	8.4	7.3
Language problem	2.6	0.1	1.8	0.3	0.2	0.2	0.3	0.1
Unavailable during fieldwork period	2.4	0.6	2.7	0.5	2.2	1.6	1.5	2.0
Non-contact	10.1	20.7	16.3	22.4	20.6	28.4	19.2	24.7
Refusal	19.3	21.6	21.1	22.1	21.8	20.7	17.7	18.6
Sick, not able	1.4	0.9	1.4	1.0	2.1	2.0	1.2	1.1
Other, no final disposition code	5.6	1.1	7.2	1.5	6.4	1.6	5.6	2.0
Total eligible sample size (in N)	2,142	1,564	2,359	1,737	2,656	1,929	2,181	1,973
Ineligibles (in N)	20	1	31	3	22	1	31	1
Total sample (in N)	2,162	1,565	2,390	1,740	2,678	1,930	2,212	1,974

Note. *p<0.05. Rounding differences can cause some columns not to add up to 100%.

These specific design choices made for the survey SIM2006 caused nonresponse among approximately 10.7% (5.7 plus 2.6 plus 2.4) of the eligible sample among Turkish compared to 3.6% in SIM2011 (Table 3.3). The same goes for the Moroccan sample which misses out on 10.5% in SIM2006 because of nonresponse due to survey design choices versus 5% of the total eligible sample in SIM2011. The difference is smaller, but similar for the Surinamese and there is hardly any difference between both samples for the Antilleans. Fourthly, there are also large and unexpected differences found in both *noncontact* rates and the final disposition code '*other, no final disposition code*' for all groups between both surveys. These outcomes are related. The main reason for the correlation is that in the SIM2006 reissue phase a high number of non-contacts were reissued, but never got

exhaustively re-contacted before fieldwork ended. For those cases, the final disposition code 'no final disposition code' was declared.
Also, in the first fieldwork phase in SIM2006 a few sampled persons never received a final disposition code, because they were not contacted the minimum number of times. The majority of these 'still not exhaustively contacted' outcomes were noncontacts up to that point. The main reasons for not following up these cases completely was due to lack of capacity (too high a workload for the interviewer) and interviewer unavailability (illness, holidays).
Finally, there are varying numbers of *ineligibles* between both surveys. The main reason for this is the pre-fieldwork check conducted by the fieldwork agency on the SIM2011 gross sample. Before the gross sample was issued to the interviewers, it was enriched with phone numbers of the sample units, if any could be found. This check also revealed ineligible sample units such as sample units that moved abroad, frame errors etc.

3.4.2 Representativity and the upper bounds of nonresponse bias among the response samples

In this section, the paradata used are the auxiliary sample frame variables. The R-indicator tells us how representative the response composition of the net sample is compared to the gross sample with respect to several specific background variables (Schouten et al. 2009). This representativity is then expressed as a single number. The variables and interaction terms used in our R-indicator model are presented in Table 3.4. The choice of variables included in the model was based on the availability of sociodemographic variables in the sample frame. No other complete frame data was available to be included in the analysis. The inclusion of interactions was based on our interest in whether or not specific difficult to survey subgroups, such as young persons living in large cities, first generation male immigrants and first generation immigrants living in large cities, were better represented using the set of design choices present in the survey design of SIM2011.

The results of the 'representativity' analysis of the response composition of the response samples show that achieving a higher response rate (RR_1) does not necessarily result in a more representative sample ($\hat{R}$)(Table 3.5).

Table 3.4
Variables and interaction terms included in the R-indicator model

Variables

Age group (15-24; 25-34; 35-44; 45-54; 55-64; 64+)

Sex (Males; Females)

Municipality size (large, midsize and small municipalities)

Immigration generation (first and second immigration generation)

Interaction terms

Age * Municipality Size
Immigration generation *Sex
Immigration generation *Municipality Size

Table 3.5
AAPOR Response rate 1 (RR_1), R-indicator ($\hat{R}$) and 95% CI, and estimates for the maximal absolute standardized bias ($\widehat{Bm}$) for each ethnic group in SIM2006 and SIM2011(in %) based on the model presented in Table 3.4

	Turkish		Moroccans		Surinamese		Antilleans	
	2006	2011	2006	2011	2006	2011	2006	2011
RR_1	52.9	52.1	43.8	48.0	40.1	41.0	46.2	44.2
$\hat{R}$	86.0	80.5	81.7	85.7	83.6	86.6	80.3	85.6
$CI_{95\%}^{\hat{R}}$	85.4	79.5	81.1	84.5	83.0	85.5	79.6	84.9
	–	–	–	–	–	–	–	–
	86.6	81.4	82.2	87.0	84.1	87.8	80.9	86.2
$\widehat{Bm}$	13.2	18.8	21.0	14.8	20.6	16.4	21.4	16.4
N[1]	2,142	1,564	2,359	1,737	2,656	1,929	2,181	1,973

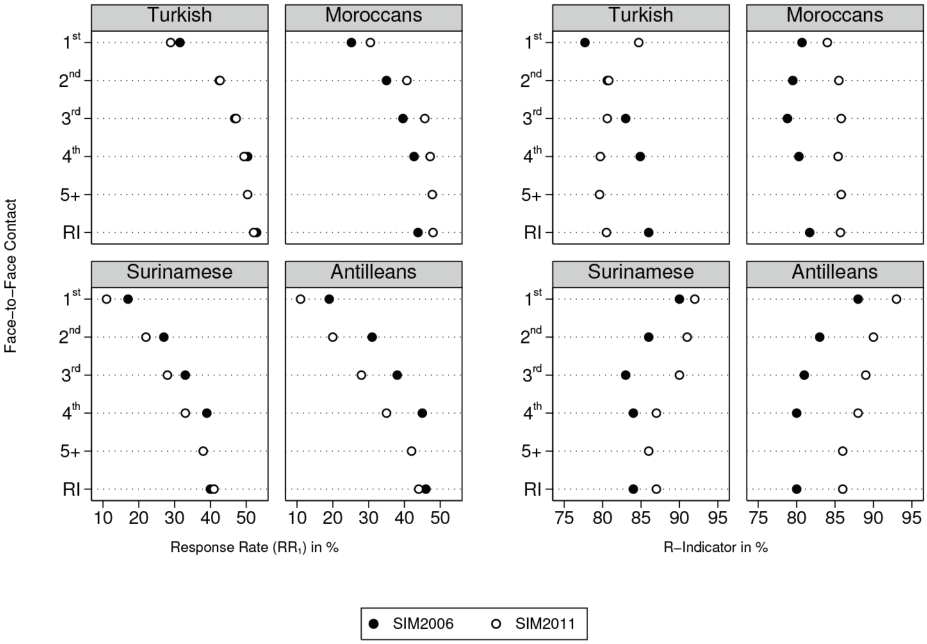
Note.[1]based on all eligible cases

The $\widehat{Bm}$ takes into account both the response rate and the response composition with respect to the variables in the model (equation 2). The combination of both indicators shows that the SIM2006 design leads to a more representative sample with a lower maximal absolute bias among the Turkish. The SIM2011 design leads to a more representative sample with lower maximal absolute standardized bias among Moroccans, Surinamese and Antilleans.

3.4.3 The evolution of sample representativity in the first and second phase of fieldwork

In this section, the paradata used are the intermediary fieldwork disposition code of the sample units and the auxiliary sample frame variables. The evolution of the sample representativity after each face-to-face contact attempt (CA) in the first phase and the effect of the reissue phase (RI) can be monitored to assess the impact or usefulness of each additional CA on the sample representativity. Of course this representativity is conditional on the previous steps, but since this is done for both surveys and separately for each of the four ethnic groups, consistent outcomes can be interpreted with more certainty.
Figure 3.1 shows the evolution of the response rate and the R-indicator for both surveys after each face-to-face CA including the RI separately for each of the four non-Western minority groups. The first contact attempt is indicated by 1, the second by 2, etc. Five or more contact attempts are indicated by 5+ and the reissue is indicated by RI. The corresponding cumulative response rate and R-indicator are presented as bars for SIM2006 and SIM2011.

Figure 3.1
The evolution of the response rate and R-indicator after each face-to-face contact attempt in both surveys separately for Turkish, Moroccans, Surinamese and Antilleans



For the Turkish sample in SIM2006 an interesting pattern is revealed. Each additional contact attempt (CA) in the SIM2006 increases the representativity of the sample. In this case a higher response rate does seem to indicate a better quality sample. Also, the targeted re-issuing was successful, improving the representativity of the sample as well.
The effect of additional CAs among Turkish in SIM2011 is somewhat different. After each additional CA during the 1st phase, the representativity decreases slightly to end a little under 80%, despite the increase in response rate after each CA. Also for this survey the reissue (RI) has a positive effect on the representativity of the response sample.
The effect during the first phase, starting at a high level, followed by a slow descent and then stabilizing is not uncommon for the evolution of the R-indicator (see for example Schouten and Cobben 2007; Schouten and Cobben 2008). As there are only few respondents, none of the subgroups based on the model used to estimate the R-indicator can be very over- or underrepresented in comparison with the other strata.
For the SIM2006 study among Moroccans, the fourth CA and the RI clearly have a positive effect on the R-indicator. This pattern is different from the SIM2011 pattern with its quick convergence. Among Moroccans in SIM2011, the additional CAs during the first phase after the second CA do not increase the R-indicator by much and the optimum seems to be just below 86%. Since there was hardly a RI among Moroccans in the SIM2011, it is clear that the impact is marginal (see also Table 3.1).
Among Surinamese, both SIM2006 and SIM2011 show the same pattern. After each of the first three CAs in the first phase, there is a significant increase in response rate, but also a drop in representativity. From the fourth CA the representativity stabilises and reaches its optimum, given the design features in this phase. The RI only increases the representativity slightly.
Both SIM2006 and SIM2011 show a similar pattern among the Antilleans. It is also very similar to the pattern among Surinamese. After each CA during the first phase the response rate increases, but the representativity decreases. It looks as if the interviewers are focusing their attention on the 'easy' respondents during the first fieldwork phase. The second phase clearly has a stabilising effect here.
Overall this analysis shows that a reissue has a positive or at least stabilizing effect on the representativity of the sample in comparison to the level of representativity at the end of phase one. This already happens with quite modest re-issuing. It seems that an extended first phase makes interviewers eventually target cases with the highest probability of success, which increases response rate, but does not (necessarily) increase the representativity of the sample. A reissue increases the representativity, probably because equal attention is again given by the new interviewer to all available sample units in the interviewers assignment.
The reissue strategies differed between SIM2006 and SIM2011. In 2006 only underrepresented subgroups got reissued to another interviewer and they received the same conditional incentive. In 2011 there was no targeted selection of underrepresented subgroups in the RI and the incentive was increased.
Despite the limited RI in both SIMs, there are some interesting differences caused by the different RI strategies (Table 3.6). It is quite clear that, in terms of response rate, the RI was much more successful in SIM2011. Also, re-issuing seems to have been more

successful among Turkish than the other ethnic groups. Almost half of the reissued cases were converted among the Turkish in 2011. However, since the increased incentives and non-targeted RI in 2011 are confounded, it cannot be determined which of the two contributed more to the increased response.

Table 3.6
The actual number of reissued sample persons and the number of successful interviews in SIM2006 and SIM2011 per ethnic group

	Actual number of sample units reissued with a final disposition code		Number of achieved interviews	
	2006	2011	2006	2011
Turkish	142	58	53	25
Moroccans	115	8	24	4
Surinamese	199	226	27	53
Antilleans	165	182	36	53
Total	621	474	140	135

The more successful RI in 2011, in terms of response rate, does not seem to result in an equal increase in representativity. In relative terms, it appears that the less successful RI in 2006 actually had a slightly larger, positive impact on the representativity

3.4.4 The evolution of the maximal absolute standardized bias in the first and second phase of fieldwork

The R-indicator shows one part of the picture, but the response rate needs to be taken into account as well in order to get an appreciation of the potential nonresponse related bias for a particular survey item. The R-indicator and the response rate are used to calculate the $\widehat{Bm}$ (see formulae 1 and 2), which serves as an estimate for the upper bound nonresponse bias on a particular survey item given the sample. Here the $\widehat{Bm}$ estimate is calculated after each contact attempt during the first phase and after the RI to show how these design features influence the upper bound nonresponse bias on a particular survey estimate. Since all these measures are part of a system of design features, the impact can only be assessed depending on the sequence preceding the measure. However, similar changes in surveys with different designs offer additional weight in evaluating the effect of each CA and a RI on the potential for nonresponse bias on survey estimates among ethnic groups.

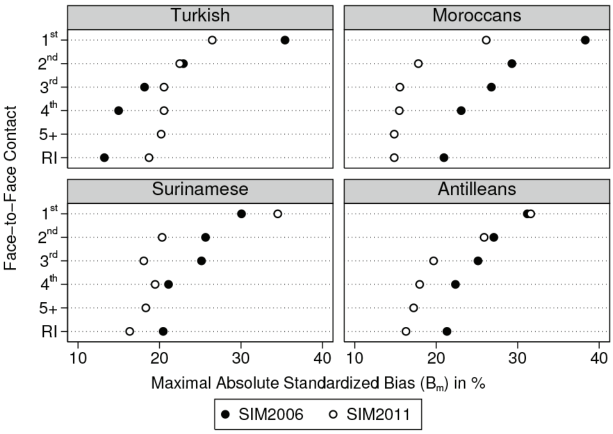
Figure 3.2 presents how the $\widehat{Bm}$ estimate in both SIM designs changes after each face-to-face CA and the RI separate for each ethnic group. The first contact attempt is indicated by 1, the second by 2 etc. Five or more CAs are indicated by 5+ and the reissue is indicated by RI.

The evolution of the $\widehat{Bm}$ estimate during the first phase of fieldwork in SIM2006 shows a slightly different picture for all four the non-Western minority groups. In this survey design, each additional CA during the first phase results in a reduced $\widehat{Bm}$ estimate and there seems to be no converging to a local minimum in the first phase. The evolution of the $\widehat{Bm}$ estimate also shows a positive effect of the RI among all groups.

Figure 3.2 shows that the call strategy of SIM2011 stabilises to a local minimum in the first phase after the third or fourth CA among all groups. The subsequent contact attempts – up to 15 in the SIM2011 during the first phase – do not result in a much reduced potential for nonresponse bias on survey estimates despite the additional response. If Figure 3.1 is compared with Figure 3.2 one can see this effect quite clearly among the Surinamese. Stopping after the third CA in the first phase and then starting the RI seems to be a more fruitful endeavour if one wants to reduce the upper bound nonresponse bias, given a fixed number of contact attempts. Also in this design the evolution of the $\widehat{Bm}$ estimate shows a positive effect of the reissue, although among the Moroccans the reissue phase was hardly implemented (see Table 3.6).

Figure 3.2

The evolution of the $\widehat{Bm}$ estimate after each face-to-face contact attempt in both surveys (SIM2006 = dark grey and SIM2011 = light grey) separately for Turkish, Moroccans, Surinamese and Antilleans

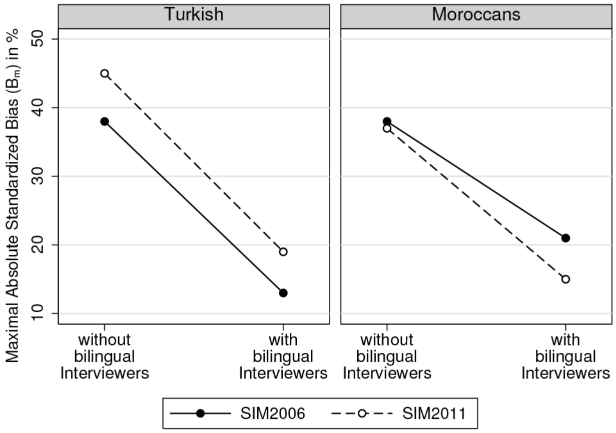


3.4.5 The effect of bilingual interviewers with a common ethnic background on the potential for nonresponse bias on survey estimates among Turkish and Moroccans

The use of bilingual face-to-face CAPI interviewers with a common ethnic background was meant to reduce nonresponse due to language problems and functional illiteracy. Both reasons can still cause response rates to drop quite significantly especially among the first generation Turkish and Moroccans in the Netherlands. This can lead to biased estimates, since it excludes a very specific group. For Surinamese and Antilleans language problems are not seen as an important cause for nonresponse since, for many, Dutch is their mother tongue.
In this section, the paradata used are the interviewer observations about the respondent's ability to read or speak Dutch and the auxiliary sample frame variables. To find out to what extent bilingual interviewers are still necessary among Turkish and Moroccans, the interviewers were asked to fill out a short questionnaire. After each successful interview, they had to answer several questions about the language in which the survey was conducted, how they assessed the respondent's proficiency in Dutch, etc. These assessments on the respondent's ability to understand Dutch were used to estimate the number of respondents that would have been missed due to language problems if no bilingual interviewers were used. In our situation, if the interview was conducted (almost) completely in their native language and the interviewer also assessed that the level of Dutch of the respondent was (very) poor, we assumed that a respondent would have been a nonrespondent due to language problems in the absence of a bilingual interviewer. This corrected response rate excluding the potential language problems, in combination with the re-estimated R-indicator enables us to re-calculate the $\widehat{Bm}$. The difference between the original and re-estimated $\widehat{Bm}$ serves as an indicator for the effect that bilingual interviewers have on the potential for language problems related nonresponse bias on survey estimates (see Figure 3.3).

There is a marked increase in the potential for nonresponse bias on survey estimates if bilingual interviewers are not used. This holds across both ethnic groups and surveys. Without bilingual interviewers, the $\widehat{Bm}$ increased about 25 percentage points on average among Turkish and about 20 percentage points among Moroccans. If the representativity of the response sample (as indicated by the R-indicator) remained equal, the increase in $\widehat{Bm}$ should have been less than the decrease in response rate (see equation 2). However, the drop in response rates was on average about 13 percentage points among Turkish and 6 percentage points among Moroccans. This suggests that the increased $\widehat{Bm}$ is largely the result of a much more unbalanced sample. This, in turn, results in an increased potential for nonresponse bias on survey estimates.

Figure 3.3
The estimated maximal absolute standardized bias ($\widehat{\mathrm{Bm}}$ in%) among the Turkish and Moroccans response sample with (B) or without (~B) the use of bilingual interviewers



3.4.6 The contribution of different subgroups to the lack of overall representativity

The lack of representativity as expressed by the R-indicator can also be partitioned into the contribution to lack of representativity of each variable included in the model to estimate the R-indicator. This is done by unconditional and conditional partial R-indicators. In this case, the larger the variation in the response propensities of a variable, the greater the contribution to the overall lack of representativity.

The unconditional and conditional variable level partial R-indicators were calculated for the variables age group, sex, municipality size and immigration generation. The unconditional partial R-indicators allow for a comparison between surveys and the conditional partial R-indicators show the unique contribution of a variable to the variability in response propensities within a survey and ethnic group, after controlling for the other variables in the model. For both indicators, the contribution of each variable to the lack of representative response is shown separately for each survey and ethnic group (Table 3.7).

Among Turkish and Moroccans, the unconditional partial R-indicator shows the largest variation in response propensities for age group. This is true for both SIM2006 and SIM2011. In the Turkish samples in both surveys, the second largest contribution comes from sex. Among the Moroccans, it comes from immigration generation, that is the imbalance of response propensities between first and second immigration generation.

Table 3.7
The unconditional and conditional variable level partial R-indicators, separate for each ethnic group and time of the survey (multiplied by 1000)

	Turkish		Moroccans		Surinamese		Antilleans	
	2006	2011	2006	2011	2006	2011	2006	2011
Unconditional								
Age group	33.2*	51.4	63.4*	20.0	42.5*	29.8	37.5	34.8
Sex	26.1*	37.6	37.9*	6.9	36.2*	4.6	48.5*	11.0
Municipality size	7.3*	26.3	22.4*	15.8	53.8*	44.9	62.9*	45.6
Immigration generation	24.7*	32.0	38.2*	17.5	12.1*	1.1	14.4*	3.4
Conditional								
Age group	28.0	60.5	51.7	23.8	41.3	31.1	36.3	37.6
Sex	27.3	36.5	40.0	5.3	35.7	5.2	50.6	11.9
Municipality size	6.4	30.5	22.3	15.3	53.4	45.5	62.4	46.8
Immigration generation	15.2	45.0	4.7	21.6	2.0	1.4	13.2	3.1
N[1]	2,142	1,564	2,359	1,737	2,656	1,929	2,181	1,973

Note * p = < 0.05 between SIM2006 and SIM2011 within ethnic groups based on confidence intervals (not included here) that were approximated using 1000 bootstrap replicates of the estimates and excluding the 25 highest en lowest estimates. [1]Based on all eligible cases.

In the Surinamese and Antillean samples in both surveys, the unconditional partial R-indicator shows municipality size as the largest contributor to the variation in the response. In both surveys the second largest contribution comes from age group among Surinamese, whereas among Antilleans, it varies per survey: in SIM2006 it is sex and in SIM2011 it is age group.

The unconditional and conditional partial R-indicators at the variable level differ in size among Turkish and Moroccans for both surveys. This means that the variables included in the model are correlated among the Turkish and Moroccan samples. In SIM2006 the contribution of immigration generation to the variation in response propensities decreases among the Turkish and Moroccans after conditioning on the other variables. Also the contribution of age group is less after conditioning, especially among the Moroccans. The conditional partial R-indicators show that, after conditioning, the two largest contributions come from age group and sex in both groups.
In SIM2011 the variables also show collinear response behaviour among the Turkish and the Moroccans. However, in this instance, the contribution of age group and immigration generation to the variation in response propensities increases after conditioning on the other variables. After conditioning, the two largest contributors among Turkish and Moroccans are age group and immigration generation.

Among Surinamese and Antilleans there is not much difference in contribution between the unconditional and the conditional partial R-indicators at the variable level in both surveys. This means there is no strong collinear response behaviour and the variables have a unique and separate impact on the representativity of the response samples among Surinamese and Antilleans.

Partial R-indicators were also estimated at the category level. These estimates can provide additional insight on how to improve on an existing survey design for a specific ethnic group by identifying under and overrepresented subgroups. The category level can also consist of categories based on an interaction of variables (Schouten et al. 2011, p. 236). We analysed a mix of two separate, single variable category level indicators and one category level indicator based on a combination of two variables (Table 3.8). Based on the conditional variable level results, the two variables that contributed the most to the variation in the response propensities were included in the interaction (Table 3.7). As a result, the interaction-category level indicators can vary between surveys and or between ethnic groups. The remaining two variable category level indicators were calculated separately for each variable.

For ease of interpretation, the category level results are shown separately for Turkish and Moroccans, on the one hand, and Surinamese and Antilleans, on the other hand. This is done because models to estimate the partial R-indicators at the category level are similar between Turkish and Moroccan in both SIM2006 and SIM2011 (Table 3.8). There is also great similarity in the models used to estimate the partial R-indicators at the category level between the Surinamese and Antillean samples in both SIM2006 and SIM2011.

Table 3.8
Overview of the models used to estimate the partial R-indicators on the category level, separate for each survey and ethnic group

Ethnic Group	SIM	Model for the estimation of the partial R-indicators
Turkish	2006	Immigration generation + Municipality size + Age group * Sex
	2011	Sex + Municipality size + Age group * Immigration generation
Moroccans	2006	Immigration generation + Municipality size + Age group * Sex
	2011	Sex + Municipality size + Age group * Immigration generation
Surinamese	2006	Sex + Immigration generation + Age group * Municipality size
	2011	Sex + Immigration generation + Age group * Municipality size
Antilleans	2006	Age group + Immigration generation + Sex * Municipality size
	2011	Sex + Immigration generation + Age group * Municipality size

The unconditional and conditional category level results show that the single largest contribution to the variation in response propensities among the Turkish 2006 sample comes from the overrepresentation of women in the age category of 35 to 44 (Table 3.9)[7].

[7] Confidence intervals were also approximated using 1000 bootstrap replicates of the estimates and excluding the 25 highest en lowest estimates and can be delivered upon request by the author.

Among the Moroccan 2006 sample there are more subgroups with a relatively large contribution (with a conditional contribution of over 20) to the variation in response propensities. These are the 15 to 34 year old men and 55 to 64 year old women, who are underrepresented, and the 35 to 54 year old women and men aged above 64, who are overrepresented.

It is interesting to note that, while the SIM2006 design sometimes leads to similar subgroups in the Turkish and Moroccan sample, such as, for instance, 15 to 24 year old males, being under (or over)represented, it also shows differences in representation of certain subgroups, such as 55 to 64 year old females, between the two samples.
As expected, the Turkish 2011 sample shows more subgroups with a large contribution to the variation in response propensities. These groups are the overrepresented women, persons living in midsize cities and first generation Turkish in the age of 15 to 24 and the underrepresented men, persons living in small municipalities and second generation Turkish in the age of 25 to 34.
Among the Moroccan 2011 sample, the underrepresented first generation immigrants aged 25 to 34 contribute the most. The complete lack of similar under and overrepresented subgroups between the Turkish and the Moroccan 2011 sample is also quite notable. Table 3.9 also shows that the Turkish and Moroccan sample did not contain any second generation immigrant above the age of 44. This was to be expected since the Turkish and Moroccan immigration only started in the mid-sixties of the last century. The immigrants were mostly men who came to the Netherlands for work. Partner reunification only started in the mid-seventies.
The Surinamese 2006 sample shows that the largest contributions to the non-representative response come from the overrepresentation of women and youngsters living in midsize and small municipalities and the underrepresentation of men and 25 to 44 year old big city dwellers (Table 3.10)[8]. In the Surinamese 2011 sample, the largest contributions come from the underrepresentation of 25 to 44 year old big city dwellers and the overrepresentation of youngsters living in small cities.

[8] Confidence intervals were approximated using 1000 bootstrap replicates of the estimates and excluding the 25 highest en lowest estimates.

Table 3.9
Unconditional and conditional partial R-indicators on category level, separate for Turkish and Moroccans for SIM2006 and SIM2011 (multiplied by 1000)

	Unconditional				Conditional			
	2006		2011		2006		2011	
	Turkish	Moroccans	Turkish	Moroccans	Turkish	Moroccans	Turkish	Moroccans
Age x Sex								
15-24 Male	−22.2	−36.3	–	–	14.0	22.4	–	–
15-24 Female	−2.5	−5.3	–	–	5.6	10.1	–	–
25-34 Male	−15.6	−28.2	–	–	13.9	29.2	–	–
25-34 Female	5.9	−0.9	–	–	7.2	12.2	–	–
35-44 Male	8.1	6.7	–	–	5.1	8.1	–	–
25-44 Female	25.2	30.8	–	–	22.4	27.3	–	–
45-54 Male	3.9	11.8	–	–	0.9	5.7	–	–
45-54 Female	14.2	32.0	–	–	10.6	24.4	–	–
55-64 Male	−7.9	15.1	–	–	11.6	7.9	–	–
55-64 Female	2.9	−27.5	–	–	1.9	21.2	–	–
64+ Male	−10.4	18.4	–	–	14.3	24.3	–	–
64+ Female	−0.8	−3.4	–	–	4.0	9.0	–	–
Age x Immig.Gen								
15-24 1G	–	–	23.4	−8.1	–	–	32.2	8.5
15-24 2G	–	–	7.9	13.6	–	–	8.1	13.8
25-34 1G	–	–	−3.9	−23.5	–	–	4.9	22.4
25-34 2G	–	–	−49.3	4.1	–	–	49.4	3.2
35-44 1G	–	–	−8.6	3.4	–	–	8.7	3.4
25-44 2G	–	–	−18.8	6.9	–	–	18.4	7.2
45-54 1G	–	–	13.8	5.4	–	–	15.2	5.1
45-54 2G	–	–	*n.a.*	*n.a.*	–	–	*n.a.*	*n.a.*
55-64 1G	–	–	14.9	−0.6	–	–	13.8	0.6
55-64 2G	–	–	*n.a.*	*n.a.*	–	–	*n.a.*	*n.a.*
64+ 1G	–	–	17.6	−0.3	–	–	17.2	1.1
64+ 2G	–	–	*n.a.*	*n.a.*	–	–	*n.a.*	*n.a.*
Municipality Size				[a]				
Large	−5.7	−0.7	−4.7	0.1	5.0	1.0	8.4	1.1
Medium	2.9	−12.9	16.5	8.2	3.5	13.4	20.3	7.9
Small	3.5	18.3	−20.0	−13.4	1.8	17.8	21.2	13.0
Immigration Generation				[b]				
1G	16.5	25.4	–	–	10.9	3.3	–	–
2G	−18.5	−28.5	–	–	10.9	3.4	–	–
Sex								
Male	–	–	−26.3	4.8	–	–	25.5	3.7
Female	–	–	26.9	−5.0	–	–	26.1	3.8
N[c]	2,142	2,359	1,564	1,737	2,142	2,359	1,564	1,737

a Large: municipality size >250,000; Medium 250,000-50,000; Small: <50000.
b 1G: first generation immigrant; 2G: second generation immigrant.
c Based on eligible cases.

Table 3.10
Unconditional and conditional partial R-indicators on category level separate for Surinamese and Antilleans for SIM2006 and SIM2011(multiplied by 1000)

	Unconditional				Conditional			
	2006		2011		2006		2011	
	Surinames	Antilleans	Surinames	Antilleans	Surinames	Antilleans	Surinames	Antilleans
Immigration Generation[b]								
1G	8.6	−8.8	−0.7	−1.8	1.4	8.8	1.0	1.9
2G	−8.5	11.4	0.9	2.9	1.3	9.9	1.0	2.4
Age x Municipality size[a]								
15-24 in L	−3.8	–	−3.5	−12.4	1.6	–	4.0	12.5
15-24 in M	20.6	–	11.1	24.2	22.8	–	10.7	23.6
15-24 in S	21.2	–	21.3	0.0	23.1	–	21.1	0.2
25-34 in L	−40.4	–	−21.6	−33.7	38.5	–	21.8	33.6
25-34 in M	−5.9	–	−5.9	−5.9	5.5	–	6.0	6.5
25-34 in S	1.0	–	9.3	−13.3	1.9	–	9.3	13.8
35-44 in L	−21.4	–	−21.1	−8.2	21.8	–	21.1	8.3
35-44 in M	8.2	–	−4.2	20.6	9.1	–	4.1	20.8
35-44 in S	13.1	–	13.4	13.6	13.7	–	13.4	11.4
45-54 in L	−3.1	–	−8.1	−3.8	5.9	–	7.8	3.5
45-54 in M	19.8	–	6.8	22.5	16.8	–	6.9	21.1
45-54 in S	21.5	–	18.5	3.8	19.8	–	18.8	4.1
55-64 in L	−6.4	–	−6.6	−12.4	5.8	–	6.5	12.1
55-64 in M	12.1	–	4.3	8.9	10.3	–	4.5	9.2
55-64 in S	12.1	–	13.3	−3.3	11.4	–	13.4	3.1
64+ in L	−1.8	–	7.4	−5.9	4.0	–	7.6	5.4
64+ in M	11.9	–	12.9	8.6	10.3	–	13.1	8.4
64+ in S	11.1	–	15.2	−0.4	11.4	–	15.3	0.4
Sex								
Male	−26.3	–	3.3	−7.6	25.9	–	3.7	8.2
Female	24.9	–	−3.2	8.0	24.6	–	3.6	8.6
Sex x Municipality size								
Male in L	–	−55.9	–	–	–	54.5	–	–
Male in M	–	−4.2	–	–	–	4.9	–	–
Male in S	–	5.6	–	–	–	4.1	–	–
Female in L	–	−14.1	–	–	–	13.2	–	–
Female in M	–	44.7	–	–	–	44.7	–	–
Female in S	–	34.7	–	–	–	33.0	–	–

Table 3.10 (continued)

	Unconditional				Conditional			
	2006		2011		2006		2011	
	Surinames	Antilleans	Surinames	Antilleans	Surinames	Antilleans	Surinames	Antilleans
Age Group								
15-24	–	12.1	–	–	–	10.5	–	–
25-34	–	–29.9	–	–	–	30.6	–	–
35-44	–	5.3	–	–	–	10.3	–	–
45-54	–	9.3	–	–	–	15.6	–	–
55-64	–	14.1	–	–	–	3.2	–	–
64+	–	–7.5	–	–	–	2.4	–	–
N[c]	2,656	1,929	1,973	2,181	2,656	1,929	1,973	2,181

a Large municipality size >250,000; Medium: 250.000-50,000; Small: <50,000.
b 1G: first generation immigrant; 2G: second generation immigrant.
c Based on eligible cases.

The category level indicators also reveal that not only the sex balance has improved in the SIM2011 sample, but also that 25 to 34 year old big city inhabitants are less underrepresented and youngsters living in midsize cities are less overrepresented compared to the SIM2006 sample. The different survey design choices made for the SIM2011 survey seem to be effective in reducing heavily over – and underrepresented subgroups.
The results for the Antillean 2006 sample show that the largest contributions to the variations in response propensities come from the underrepresentation of men living in the big cities and 25 to 34 year old persons and the overrepresentation of women living in midsize and small municipalities (Table 3.10).
The 2011 sample shows the largest contribution coming from the underrepresented big city dwellers aged 25 to 34 and the overrepresented youngsters and persons between the ages 35 to 54 living in midsize cities. It is also interesting to see that the SIM2011 design leads to quite a few differences in the over and underrepresented subgroups among Antilleans compared to the Surinamese.
The results of the variable and category level partial R-indicators analysis have shown which groups are over- and underrepresented among the different ethnic groups in the SIM2006 and SIM2011 survey. The analyses have shown that different subgroups are under and overrepresented across the various ethnic groups and surveys. This means that the survey design and the characteristics of the population under study cannot be viewed as separate entities that affect the likelihood of response, but should be viewed as an interactive system. For instance, if one takes the SIM2011 design as a basis to conduct another survey among the same four ethnic groups, varying targeted data collection strategies should be developed depending on the ethnic group, but these strategies for the same ethnic groups would be different if the SIM2006 were to be used as a basis. In addition, when developing group dependent data collection strategies, one should

not only look at the characteristics of the underrepresented subgroup, but also at their cause for nonresponse.

For example, to increase the representativity among a sample of Moroccans using the SIM2011 design, it is likely that one needs to increase the response among first generation immigrants in the age of 25 to 34. The characteristics of the subgroup tell us that these are people who have come to a new country and could be unfamiliar with the Dutch culture or language. They could have come to the Netherlands to get married or to find work. In order to improve the probability of response among this subgroup one can choose different methods, such as using a different data collection mode (i.e., CAWI in case the potential respondent is away during interviewer working hours or in case the potential respondent is unwilling to communicate with an interviewer), increasing the number of contact attempts, using higher incentives or sex matching the interviewer to the potential respondent. The relatively high noncontact rate among this subgroup would suggest that a sex match or increased incentives might not be the preferred tailored approach, but that another data collection mode or increasing the number of contact attempts might be more applicable.

Targeting a different subgroup using a different method would have been appropriate among the Turkish in the 2011 sample. In that case, 25 to 34 year old second generation immigrants were underrepresented and the refusal rate was relatively high. In the 2006 sample various other subgroups were underrepresented among the Moroccans and also the cause for nonresponse differed between the various subgroups.

3.5 Conclusion and Discussion

Surveying among non-Western minorities continues to be difficult, but focussing on other indicators instead of only the response rate as measures of quality might prove insightful in the pursuit of a more representative sample among non-Western minorities – or other populations for that matter. In this paper we focused on how different survey design choices affect the composition of the response sample and how this might relate to the occurrence of nonresponse bias on survey estimates in surveys among non-Western minorities in the Netherlands.

It is important to know about the survey related characteristics of the population of interest when designing a survey. Each design choice can potentially lead to the exclusion of target population members, therefore the more aware one is of these survey related characteristics, the more informed the tradeoff decision. Fieldwork disposition codes show that basic survey design decisions, such as the intended length and timing of fieldwork, can result in increased nonresponse among ethnic groups in the Netherlands, because of higher mobility among non-Western minorities and unavailability due to long holidays in the country of origin. Especially the use of a long fieldwork period increases the likelihood of nonresponse due to the fact a sampled person may have moved.

The results from the R-indicator analysis show that different survey designs lead to different levels of representativity of the response samples among non-Western minority groups in the Netherlands. Furthermore, the level of representativity seems to be uncorrelated with the response rate when the difference between response rates is significant. A higher response rate under these conditions does not necessarily result in a more representative sample.

The estimated maximal absolute standardized bias, where the R-indicator is used in conjunction with the response rate, shows that the potential for nonresponse bias on substantive outcomes can be quite substantial. This result raises concerns on the validity of results concerning non-Western minorities obtained from non-Western respondents in general population surveys, because less extensive measures are usually undertaken to reach non-Western minority groups.

All in all, the results have shown that it is possible, given the right survey design, to combat declining response rates and increase the quality of response samples in surveys among hard-to-reach populations, such as non-Western minorities This is even possible despite the potentially harmful effect of a more populist discourse on migrants on the willingness to participate in the Netherlands.

The impact of several survey design choices on the potential for nonresponse bias on survey estimates was also analysed in more detail. The results showed that the optimal number of face-to-face contact attempts in a multi-phase approach of non-Western minorities in the Netherlands is about three to four in the first phase. More contact attempts made in the first phase by the same interviewer do increase the response, but do not decrease the potential for nonresponse bias on survey estimates. Limiting the number of contact attempts to a maximum of four during the first phase before moving to a reissue phase in which other design features can be introduced can potentially result in significant time and/or financial gain.

The reissue phase, in which non-responding sampled persons were contacted by another interviewer, reduced the potential for nonresponse bias on survey estimates and increased the representativity of the response sample composition among all non-Western minorities samples. This is despite the fact of some serious shortcomings in the execution of the reissue phase among all samples in the current study. All samples used in this study had a far more limited reissue phase than initially planned. Let this serve as a reminder to always plan enough time to conduct a reissue phase and to ensure the availability of enough interviewers. Even so, the analysis showed that even a quite modest reissue had a positive effect on the sample composition. It is self-evident that if the reissue had been fielded as intended, the response would have been higher and based on this analysis, the nonresponse bias of the survey estimates should also have been reduced.

This study suggests that an increased conditional nonmonetary incentive during the reissue phase does not necessarily result in a larger decrease of potential nonresponse bias among non-Western minority groups compared to keeping the conditional incentives at the same level. However, the effect of an increased incentive is most likely confounded with the way the reissue in the SIM2011 design was designed. In the 2011design, reissued persons did not necessarily belong to underrepresented subgroups. This is different from the targeted reissue that was applied in the SIM2006 design. From a cost perspective and bias reduction point of view, it may be better not to use an increased conditional nonmonetary incentive and reissue all non-responding sampled persons in the second phase of surveys among non-Western minorities, but rather to target underrepresented subgroups. Of course, one needs to be careful and realize that the maximal absolute standardized bias is only an indicator for nonresponse bias on survey estimates. Also, when targeting underrepresented subgroups a different payment scheme for interviewers might be in order to keep them motivated.

Interviewers with a common ethnic background remain of great importance in order to reach a balanced or representative sample among non-Western minorities. Obviously, the use of bilingual interviewers with a common ethnic background reduces the nonresponse due to language problems and also the potential for nonresponse bias on survey estimates. Especially among ethnic groups with known language problems, the possibility of quite severe nonresponse bias on survey estimates exists if one does not use bilingual interviewers. Reducing the potential for nonresponse bias on survey estimates by minimizing language problems is only one of the benefits of using interviewers with a common ethnic background. The results of the partial R-indicators also suggest that other difficult subgroups without any known language problems, such as young second generation Moroccan immigrants or Antillean men living in large cities, are also better represented and sometimes even overrepresented in the response samples among non-Western minorities when interviewers with a common ethnic background are more extensively used. Of course, the effectiveness of interviewers with a common ethnic background is evaluated here in terms of potential for nonresponse bias on survey estimates, but this is only part of the survey cycle. Interviewers with a common ethnic background may also have a greater influence on the way respondent answers to survey questions, compared to interviewers without a common ethnic background, which can lead to increased measurement bias. One should be aware of this trade-off.

When it comes to evaluating the effect of separate response-enhancing measures in surveys it is important to note that in many circumstances analysis methods, such as logistic regression, give biased results because of non-random allocation of sample units to ‘treatments’. Brehm (1993, p. 128-130) also sees this inherent problem in applying the continuum of resistance to reluctance. He combined a continuum of resistance with respect to accessibility and to cooperation in his approach to modelling the survey process, in which even more administrative measures (more calls, sending a letter to try and persuade reluctant sample persons, trying to convert a refusal) would increase survey participation. The difficulty he found with this model is that persuasion letters are only sent to reluctant respondents, and therefore seem to have a negative effect (as reluctant respondents more often turn into final refusers and no persuasion letters are sent to respondents who cooperate instantaneously). As he remarks in a footnote (p. 130): ‘If one’s interest lies in how effective these techniques are [...], the persuasion letters and refusal conversions would have to be randomly assigned treatments, not treatments assigned on the basis of an initial refusal.’

It is also important to realize that one size does not fit all when designing a survey among different non-Western minority groups. The results of the unconditional and conditional partial R-indicators showed that there are significant differences in under and overrepresented subgroups depending on survey design and ethnic group. This is important to keep in mind when one is trying to assess whether non-Western minorities are well represented in a general population survey. An underrepresented sociodemographic subgroup among one ethnic group might be cancelled out by the overrepresentation of the same subgroup among another ethnic group. This will lead to a biased result if the two subgroups have different views or attitudes based on their culture or socioeconomic status as an ethnic group.

Fieldwork strategies can be improved and tailored to reach hard-to-reach subgroups. The partial R-indicators in conjunction with the final fieldwork disposition codes provide a wealth of information for improving the representativity of a survey among different non-Western minority groups. They can tell us not only who to target, but also how we should target them.

Finally, the approaches used in the analysis provide us with additional insight on the quality of the response sample and on the occurrence of nonresponse bias at survey item level. However, one should keep in mind that these approaches use the information available at survey level to assess the potential for nonresponse bias at item level. However, nonresponse bias is item specific and not survey specific (Groves and Peytcheva 2008). The predictive value of fieldwork disposition codes or the R-indicator in conjunction with the maximal absolute standardized bias based on auxiliary variables can be quite limited when estimating the actual size of the nonresponse bias, but the combination of these approaches will tell us more about the potential for nonresponse bias on survey estimates than the response rate alone.

4 The Impact of Face-to-Face versus Sequential Mixed-Mode Designs on the Possibility of Nonresponse Bias in Surveys among non-Western minorities in the Netherlands

In this chapter we compare the quality of realized samples based on a single-mode CAPI survey design with the quality of realized 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 realized samples and the estimated potential for nonresponse bias in survey estimates based on auxiliary variables and the response rate. This chapter also investigates if these designs systematically enhance response rates differently among various sociodemographic subgroups based on auxiliary variables. Furthermore, 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 realized samples and show more potential for nonresponse bias in survey estimates. In addition, the designs lead to systematic differences in the level of representative response 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.[1]

4.1 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] This chapter has been published as Kappelhof, J.W.S. (2015). The Impact of Face-to-Face versus Sequential Mixed-mode Designs on the Possibility of Nonresponse Bias in Surveys among non-Western minorities in the Netherlands. *Journal of Official Statistics*, Vol. 31, No. 1, pp. 1-31.

probability of completing an interview, especially among nonwestern 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 chapter 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. wether these designs systematically enhance response rates differently among various sociodemographic 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 chapter, 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 chapter 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.

4.2 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 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).

4.3 Data and Methods

4.3.1 Data

The Dutch Survey on the Integration of Minorities (SIM) sets out to measure the socio-economic 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 study. 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 study.

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 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 4.1).

Table 4.1
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

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 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: €7.0 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 €12.0. 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 population (Kappelhof 2010). For the SIM survey, auxiliary information about ethnicity, age, gender, municipality and status as first- or second-generation immigrants was available for the sample frame data for all sampled persons. This allowed for a tailored approach of 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.

4.3.2 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 is 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 2008; Schouten et al. 2009; Wagner 2010) or the weakest assumption, *missing not at random* (MNAR 2010) (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 representativity or 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, p. 154; Little and Rubin 2002).

Schouten et al. (2009, p. 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 ($\hat{\hat{\rho}}$), the R-indicator $\hat{R}\left(\hat{\rho}\right)$, the estimated standard deviation of the survey item $\hat{S}(y)$ and the maximal absolute bias $\widehat{B_m}(\hat{\rho}, y)$.

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

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):

$$\widehat{B'_m}\left(\hat{\rho}, y\right) = \frac{\left(1 - \hat{R}\left(\hat{\rho}\right)\right)}{2\hat{\hat{\rho}}} \qquad (2)$$

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 $\widehat{B_m}(\hat{\rho}, y)$ and the FMI approach differ in the way they estimate how nonresponse bias can impact the survey estimate. For instance, the $\widehat{B_m}(\hat{\rho}, y)$ presented in Equations (1) and (2) is an estimate of the upper bound non-response 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 $\widehat{B_m}(\hat{\rho}, y)$ 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 $\widehat{B_m}(\hat{\rho}, y)$.

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, p. 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, p. 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, p. 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, p. 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.

4.4 Results on the comparison of single and mixed-mode designs among ethnic minorities

4.4.1 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, p. 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 4.2, 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 $\widehat{Bm}$ takes into account both the response rate and the response composition with respect to the variables in the model (Eq. 2). The $\widehat{Bm}$ shows similar results to the R-indicator. The single-mode *response* samples all result in lower $\widehat{Bm}$ estimates than their mixed-mode counterparts.

Table 4.2
Response rate (RR_1), R-indicator ($\hat{\mathrm{R}}$), 95%-confidence interval R-indicator ($\widehat{R}^{CI}_{0.95}$, maximal absolute standardized bias $\widehat{\mathrm{Bm}}$ 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 (%)	$\hat{\mathrm{R}}$ (%)	$\widehat{R}^{CI}_{0.95}$ (%)	$\widehat{Bm}$ (%)	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	978
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.

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 $\widehat{Bm}$ estimate, the SM design always leads to lower estimates for the upper bound nonresponse bias than the MM design-based estimates.

4.4.2 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 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 on 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 (Appendix 4.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 4.3 presents the summary results of the analysis and the actual FMI estimates are shown in Appendix 4.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.

Table 4.3

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.

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 than the MM design (see for instance Table 4.2). We could therefore have expected that within an 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 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 4.3, 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 ($\widehat{Bm}$) and the mean of the 16 fraction of missing information estimates ($\widehat{FMI}$)

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 $\widehat{Bm}$ with the eight outcomes of the $\widehat{FMI}$ (plus standard deviation) to check whether or not they lead to similar conclusions (Table 4.4). We have chosen to use the $\widehat{FMI}$ 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.

Table 4.4
The estimated maximal absolute standardized bias ($\widehat{Bm}$), the mean and standard deviation of the 16 fraction of missing information estimates ($\widehat{FMI}$) separately for SM and MM and ethnic group

	Turkish		Moroccans		Surinamese		Antilleans	
	SM	MM	SM	MM	SM	MM	SM	MM
$\widehat{FMI}$	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)
$\widehat{Bm}$	18.8	21.4	14.8	23.4	16.4	22.4	16.4	23.4

The results differ somewhat if we compare both survey designs across all ethnic groups (Table 4.4). For instance, the lowest $\widehat{Bm}$ does not correspond with the lowest $\widehat{FMI}$. Also, the four lowest $\widehat{Bm}$s estimates all come from SM *response* samples, whereas this is only true for three out of the four lowest values of the $\widehat{FMI}$. However, the results are quite similar if we compare the indicators within an ethnic group. Within each ethnic group, both $\widehat{Bm}$ and $\widehat{FMI}$ are lower when they are based on the SM data than on the MM data. This result makes sense because, while the $\widehat{Bm}$ 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 same ethnic population. Still, we would gather that both estimates lead to the conclusion that the SM design outperforms the MM design.

4.4.3 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 4.5 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 4.6.

Tables 4.5 and 4.6 also contain the main findings of the single-mode survey design, separately for each ethnic group. Appendix 4.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 overrepresentative response 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 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 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 generation* show a strong collinear response behavior among the Turkish and the Moroccans (see unconditional and conditional partial R-indicators in Appendix 4.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.

Table 4.5
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 'o' = representative. +, o and - are based on whether or not zero is included in the approximated confidence interval.

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 a 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 (Appendix 4.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 underrepresented 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 4.6). 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.

Table 4.6
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; '-' = under-represented; 'o' = representative. +, o and - are based on whether or not zero is included in the approximated confidence interval.

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.

4.4.4 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 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 4.7 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.

Table 4.7

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

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 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.

4.5 Conclusion and discussion

In this chapter 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 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 surveys-designs among non-Western minorities. The first approach was the representativity indicator (R-indicator) and the maximal absolute standardized bias ($\widehat{Bm}$) 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 $\widehat{Bm}$ 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 recommended 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, $\widehat{Bm}$ 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 reissue 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 reissue. In the reissue, 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 (See Chapter 3).

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.

Appendices

Appendix 4.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 4.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
$\mathrm{FMI}_{\text{Ethnic self}}$	44.5	51.4	46.0	51.6	51.2	69.9	44.2	60.1
$\mathrm{FMI}_{\text{Employment}}$	43.9	41.2	48.2	48.3	49.8	66.0	42.9	56.9
$\mathrm{FMI}_{\text{Religious}}$	43.8	52.4	48.2	45.6	54.4	68.5	41.0	60.2
$\mathrm{FMI}_{\text{Happiness}}$	50.7	56.1	51.3	58.1	56.7	75.0	47.3	65.6
$\mathrm{FMI}_{\text{Self}-\text{identification}}$	53.9	63.0	58.0	56.0	56.9	74.2	56.4	57.0
$\mathrm{FMI}_{\text{Health}}$	41.7	53.2	48.4	55.5	55.8	74.1	47.8	65.3
$\mathrm{FMI}_{\text{House}}$	45.8	49.6	53.4	50.5	53.0	65.4	47.9	57.5
$\mathrm{FMI}_{\text{Discriminatoin self}}$	45.3	51.2	51.7	55.9	50.9	74.7	61.5	63.5
$\mathrm{FMI}_{\text{Opportunities}}$	47.1	56.8	55.7	57.7	56.2	72.6	60.3	61.6
$\mathrm{FMI}_{\text{Children}}$	36.7	40.9	44.3	43.3	45.1	59.6	42.9	57.8
$\mathrm{FMI}_{\text{Satisfaction \_ Society}}$	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
$\mathrm{FMI}_{\text{Internet}}$	42.7	44.1	48.1	52.3	48.0	70.6	50.5	57.5
$\mathrm{FMI}_{\text{Man \_ finance}}$	45.1	52.6	52.4	59.8	55.7	77.0	53.5	62.6
$\mathrm{FMI}_{\text{Speak \_ Dutch}}$	36.2	51.1	40.1	56.1	51.4	58.5	48.4	61.8
$\mathrm{FMI}_{\text{Sports \_ frequency}}$	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
$\mathrm{NR}_{\text{Self \_ identication}}^{1}$	48.0	46.0	52.5	49.3	59.8	57.9	56.7	56.2
$\mathrm{NR}_{\text{House}}^{1}$	48.4	46.3	53.3	48.8	59.1	56.9	56.0	56.1
$\mathrm{NR}_{\text{Discrimination \_ self}}^{1}$	48.0	46.2	53.1	49.0	59.1	57.1	56.2	56.3
$\mathrm{NR}_{\text{Opportunities}}^{1}$	48.1	46.3	52.7	49.3	59.6	57.9	56.6	56.9
$\mathrm{NR}_{\text{Satisfied \_ Society}}^{1}$	48.2	45.7	52.6	48.6	59.2	57.6	55.9	55.8
$\mathrm{NR}_{\text{MD}}^{1}$	49.2	47.2	54.1	51.6	59.3	58.6	55.9	57.0
$\mathrm{NR}_{\text{Man \_ finance}}^{1}$	48.1	45.6	52.6	49.3	59.1	57.4	56.1	55.9
N	1,564	978	1,737	1,086	1,929	1,203	1,973	1,231

Note: ^{1} is corrected for item nonresponse.

Appendix 4.C

Table 4. C1: The unconditional variable and category level partial R-indicators (multiplied by 1000) and response rate (RR_1), representativity indicator ($\hat{R}$) and maximal absolute standardized bias ($\widehat{Bm}$ in %, separate for each ethnic group and survey design stage

	Turkish				Moroccans			
	MM	MM	MM	SM	MM	MM	MM	SM
	WEB	WEB + Cati			WEB	WEB + Cati		
Unconditional								
Age group	60.8	53.1*	79.0	51.4*	60.7	60.4*	87.7	20.0*
15-24	43.9	39.5	61.6	21.1	35.4	37.9	70.0	8.5
25-34	7.0	-5.9	-44.0	-36.0	17.2	12.7	-11.0	-16.6
35-44	-9.2	2.9	-1.3	-13.1	-9.5	-13.5	-21.6	4.7
45-54	-23.1	-18.6	-22.3	13.8	-20.6	-17.6	-7.3	5.4
55-64	-16.5	-19.5	2.8	14.9	-21.1	-12.6	-29.7	-0.5
64+	-28.8	-22.1	-1.0	17.6	-34.2	-37.4	-35.6	-0.0
Sex	4.1*	16.0	18.3	37.6*	37.3	37.5*	17.9	6.9*
Male	2.8	-11.0	-12.7	-26.2	-26.1	-26.3	-12.5	4.8
Female	-3.0	11.5	13.2	26.9	26.6	26.8	12.8	-5.0
Municipality	4.0*	16.9	17.0	26.3*	18.9	17.6*	51.2	15.8*
Large	-2.2	-13.1	-13.7	-4.7	-13.7	-11.7	-31.6	0.1
Medium	-0.1	10.6	8.7	16.5	11.8	5.3	10.3	8.2
Small	3.3	13.3	5.0	-20.0	5.1	12.1	38.9	-13.5
Immigration generation	52.0	44.6*	33.0	32.0*	51.6	54.2*	65.4	17.5*
1G	-29.7	-25.4	-18.9	18.3	-29.8	-31.3	-37.8	-9.7
2G	42.7	36.6	27.1	-26.2	42.1	44.2	53.3	14.6
RR_1	21.4	25.3	54.5	52.1	22.7	24.0	51.7	48.0
$\hat{R}$	84.6	84.9	76.8	80.5	83.0	82.0	75.8	85.7
$\widehat{Bm}$	36.0	29.8	21.4	18.8	37.4	37.5	23.4	14.8
N[1]	978	978	978	1,564	1,086	1,086	1,086	1,737

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.

Surinamese				Antilleans			
MM	MM	MM	SM	MM	MM	MM	SM
WEB	WEB + Cati			WEB	WEB + Cati		
34.0	27.9*	48.9	29.8*	37.8	35.4	34.8	34.8
23.0	12.0	18.2	11.4	18.5	8.5	18.5	9.3
-12.2	-20.4	-34.4	-16.1	-6.2	-12.3	-26.8	-28.7
3.8	5.6	-0.8	-12.2	9.2	11.6	11.4	9.5
0.8	1.1	5.0	4.8	-21.3	-18.5	-3.4	14.3
-21.3	-8.4	-8.9	2.1	4.8	19.0	-2.0	-2.2
-1.7	10.6	27.6	17.9	-21.9	-13.6	-2.4	1.8
2.2*	18.4	27.9	4.6*	27.9	33.8*	17.4	11.0*
-1.6	-13.3	-20.2	3.2	-19.9	-24.0	-12.4	-7.6
1.5	12.7	19.2	-3.1	19.6	23.7	12.2	8.0
40.6*	49.1*	31.5	44.9*	39.1	37.8*	80.7	45.6*
-25.6	-32.7	-16.5	-24.5	-21.3	-23.3	-65.0	-33.1
31.6	36.1	25.3	7.9	-1.3	2.2	43.9	31.0
-0.7	6.1	-8.7	36.8	32.7	29.6	19.0	-4.4
41.8*	25.0*	14.1	1.1*	63.7*	48.1*	18.2	3.4*
-25.3	-15.1	-8.5	-0.6	-33.6	-25.3	-9.6	-1.8
33.3	19.9	11.2	0.9	54.1	40.9	15.4	2.9
20.9	26.9	43.1	41.0	21.6	26.2	44.4	44.2
83.9	82.1	80.7	86.6	80.4	81.4	79.1	85.6
38.5	33.3	22.4	16.4	45.4	35.5	23.4	16.4
1,203	1,203	1,203	1,929	1,231	1,231	1,231	1,973

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

	Turkish				Moroccans			
	MM	MM	MM	SM	MM	MM	MM	SM
	WEB	WEB + Cati			WEB	WEB + Cati		
Conditional								
Age group	35.6	33.9	71.8	60.5	36.2	34.0	60.6	23.8
15-24	19.4	17.8	46.7	39.7	11.3	10.6	39.7	6.2
25-34	15.5	15.6	50.2	36.3	17.8	13.1	31.4	18.3
35-44	10.4	14.5	13.0	22.9	8.3	6.3	12.5	9.8
45-54	10.2	7.5	14.0	6.5	4.5	2.1	11.8	9.2
55-64	6.6	12.0	7.7	7.8	13.6	4.1	17.0	2.9
64+	20.0	13.5	5.9	11.8	24.3	26.5	22.8	2.1
Sex	2.6	17.5	20.2	36.5	33.9	34.9	15.6	5.3
Male	1.8	12.0	14.0	25.4	24.0	23.7	11.0	3.7
Female	1.9	12.6	14.5	26.1	23.9	23.6	11.2	3.8
Municipality	5.0	11.8	12.9	30.5	22.9	18.8	51.5	15.3
Large	1.6	9.5	10.2	8.3	15.8	13.5	34.2	1.1
Medium	3.3	6.4	7.1	20.3	16.4	9.1	15.6	7.9
Small	3.3	2.8	3.1	21.1	2.2	9.3	35.2	13.0
Immigration generation	16.5	16.0	2.0	45.0	17.0	21.0	9.2	21.6
1G	11.6	11.0	1.3	32.2	12.2	15.2	6.5	15.4
2G	11.7	11.6	1.5	31.4	11.8	14.5	6.5	15.2

Surinamese				Antilleans			
MM	MM	MM	SM	MM	MM	MM	SM
WEB	WEB + Cati			WEB	WEB + Cati		
32.7	35.8	52.2	31.1	27.8	34.0	29.6	37.6
6.5	7.2	14.1	11.6	2.3	4.9	15.4	10.5
22.1	27.5	38.4	15.6	5.4	10.2	22.4	30.6
12.6	12.6	6.0	13.9	13.3	15.5	10.7	10.3
14.0	11.0	10.5	4.9	11.3	11.8	1.9	15.6
12.8	5.2	6.9	3.0	12.2	23.0	3.3	3.2
4.0	12.9	29.2	19.0	16.9	11.1	3.6	2.4
4.0	19.7	29.0	5.2	31.5	34.9	17.6	11.9
2.9	14.2	21.0	3.7	22.1	24.7	12.5	8.1
2.8	13.6	20.0	3.6	22.4	24.7	12.4	8.6
38.0	46.6	30.5	45.5	33.7	31.4	77.8	46.8
23.7	30.8	15.4	25.3	17.6	19.1	62.2	32.8
29.5	34.4	24.4	9.1	5.0	4.0	43.2	32.5
3.3	6.0	9.7	36.7	28.3	24.6	17.8	7.7
37.2	29.9	20.6	1.4	54.1	45.1	7.3	3.1
24.5	19.7	14.1	1.0	33.2	27.5	4.4	1.9
27.9	22.4	15.0	1.0	42.7	35.8	5.8	2.4

5 Estimating the impact of measurement differences introduced by efforts to reach a balanced response among non-Western minorities in the Netherlands

This chapter investigates the impact of different modes and tailor-made response-enhancing measures (TMREM) – such as bilingual interviewers with a shared ethnic background and translated questionnaires – on the measurement of substantive variables in surveys among minority ethnic groups in the Netherlands. The data used in this study come from a large scale survey design experiment among the four largest non-Western minority ethnic groups in the Netherlands comparing single-mode CAPI and sequential -CAWI-CATI-CAPI- mixed-mode. The number and intensity of the TMREM varied among the four ethnic groups. The results show that measurement effects occur among all ethnic groups and are the result of a combination of mode-effects and TMREM. Measurement effects occur more often on sociocultural questions, but also, on occasion, on more sociostructural or background questions.[1]

5.1 Introduction

Non-western minority ethnic groups in the Netherlands are difficult to survey because of cultural differences, language barriers, sociodemographic characteristics and a high mobility (Feskens et al. 2010; Kappelhof 2010; Schmeets and van der Bie 2005). As a result, ethnic minorities are often underrepresented in surveys (Feskens et al. 2006; Schmeets and van der Bie 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). This is why in the Netherlands separate surveys among the main minority ethnic groups (i.e., non-Western minorities) continue to be necessary.

One important part of any survey design is the data collection mode (face-to-face, telephone, web or mail). These modes differ in costs, but also in response rates (Hox and De Leeuw 1994; Lozar et al. 2008). Among non-western minorities, telephone, web and mail questionnaires all lead to increased nonresponse due to either 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 et al. 2010; Gijsberts and Iedema 2011; Korte and Dagevos 2011; Schothorst 2002; Van Ingen et al. 2007).

To reduce nonresponse due to language barriers or cultural differences, it is often necessary to make use of Tailor-Made Response-enhancing Measures (TMREM), (Feskens et al. 2010; Kappelhof 2010; Kemper 1998). Examples of TMREM are the use of translated questionnaires, bilingual interviewers and interviewers with a shared ethnic background. However, the combination of a face-to-face survey and TMREM among non-Western

[1] This chapter has been conditionally accepted as Kappelhof, J.W.S. and De Leeuw, E.D. Estimating the impact of measurement differences introduced by efforts to reach a balanced response among non-western minorities. Sociological Methods and Research.

minorities is becoming more and more costly. As a consequence, despite the known limitations of other modes of data collection, there is a strong demand for exploring the possibility of employing less expensive methods of data collection and TMREM among non-Western minorities. Preferably, this has to happen without any substantial loss of survey data quality (Biemer and Lyberg 2003). One possible way of reducing costs and dealing with the additional nonresponse brought about by the different modes is the use of a sequential mixed-mode survey (De Leeuw 2005; Tourangeau 2013). However, the use of mixed-mode surveys is known to enhance the risk of measurement bias and variability of survey estimates, mostly because it is difficult to disentangle mode and selection effects (De Leeuw et al. 2008; Dillman and Christian 2005; Voogt and Saris 2005). The combination of a sequential mixed-mode survey and the use of TMREM may further increase the measurement variability. This trade-off between reducing costs and dealing with increased measurement error or variability may even outweigh the financial gains of using a sequential mixed-mode survey. An important question is therefore whether the combination of a sequential mixed-mode survey and TMREM among non-western minorities in the Netherlands is an acceptable alternative to a single-mode face-to-face survey with TMREM in terms of measurement error or variability.

Our research aim is to investigate the impact of different modes in conjunction with TMREM on the measurement of substantive variables in surveys among non-Western minorities in the Netherlands. To what extent does the use of different modes together with TMREM elicit measurement differences? In order to assess measurement effects, we will use a recently developed technique for disentangling mode and selection effects (Vannieuwenhuyze et al. 2010; Vannieuwenhuyze et al. 2012).

The data used in this study come from a large scale survey design experiment. Statistics Netherlands drew two random samples of named individuals 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 data collection was done by Gfk Netherlands and the fieldwork for both survey conditions was conducted simultaneously and lasted from November 2010 until June 2011.

This article starts with an overview of the main challenges to the measurement of substantive variables in the context of cross-cultural research when using a sequential mixed-mode approach. The experiment and the method used to answer our research question are described in more detail in the third section, followed by the results and the subsequent conclusion and discussion.

5.2 The impact of mode effects, cultural differences and/or language barriers on measurement error in cross-cultural mixed-mode survey research

The same survey question can yield different results depending on the mode used to collect the data (Couper et al. 2004; De Leeuw 1992; Tourangeau and Yan 2007). In mixed-mode survey research it is not always easy to assess the impact of mode on measurement error, as different modes may have different levels of interpenetration among the target population, vary in response rates and/or have a different impact on measurement and its associated measurement error (Feskens et al. 2010; Tourangeau 2013). When a sequential mixed-mode survey is used to collect the data, it is nearly impossible to disentangle mode effects from selection effects. Mode differences in the measurement of substantive variables can be caused by either mode or selection or both (De Leeuw 2005; Voogt and Saris 2005).

In the context of cross-cultural survey research, measurement differences may not only be introduced by the use of different data collection methods, but also by the respondent-interviewer interaction or by the race/ethnicity of the interviewer. Research has shown that the ethnicity of the interviewer has an effect on the way a respondent answers to a survey question (Anderson et al. 1988; Davis 1997; Finkel et al. 1991). Especially the match between race of the interviewer and that of respondent influences answers given on culturally sensitive questions (Van't Land 2000).

In cross-cultural survey research, difficulties in understanding the main survey language among certain cultural groups or subgroups can increase measurement error. For example, if one or more translations of the survey are necessary it can increase measurement error if the translations are not equivalent (Harkness and Schoua-Glusberg 1998; Harkness et al. 2003; Harkness et al. 2004). This can be even more pronounced if the language is a spoken only language, such as the Berber language, which is the main language of many Dutch of Moroccan origin living in the Netherlands. Furthermore, there is a possibility of increased measurement error if there are cultural differences in understanding the concept being measured or the question aiming to measure it (Hui and Triandis 1983; Van de Vijver 2003; Liang et al. 1987). For a comprehensive overview of the difficulties and best practices for conducting comparative survey research across cultures and countries, we refer to the Cross-Cultural Survey Guidelines (CCSG) (Survey Research Center 2010) or Harkness et al., (2010).

5.3 Data and Methods

5.3.1 Data

The Dutch Survey on the Integration of Minorities (SIM) sets out to measure the socio-economic 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 among each of the four largest non-Western minority populations: Dutch of Turkish, Moroccan, Surinamese, and Antillean[2] descent.

In the Netherlands, the official definition as used by Statistics Netherlands, of Dutch of Turkish, Moroccan, Surinamese, and Antillean descent includes persons that were either born in Turkey, Morocco, Surinam or the Dutch Antilles including Aruba or have at least

[2] Including Aruba.

one parent who was born there. In case 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. In 2013, these four ethnic groups make up about two-thirds of the total non-Western population which amounts to approximately 7% of the total 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.

For this experiment, Statistics Netherlands drew eight samples of named individuals from the population register: two random samples were drawn from each of four mutually exclusive population strata: Turkish, Moroccan, Surinamese, and Antilleans living in the Netherlands, aged 5 years and above. The same stratified two stage probability sampling design in all four population strata was used to draw the samples. The samples were stratified by municipality size (i.e., municipalities of more than 250k inhabitants, municipalities between 250k and 50k, and municipalities with less than 50k inhabitants). The first stratum consisted of four municipalities; in each of these municipalities a number of named individuals was drawn proportional to the size (PPS) of the municipality. In the other two strata municipalities were drawn as PSUs (PPS) and within each PSU a cluster of named individuals was drawn. 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 nonrespondents were reissued to another CAPI interviewer who had to conduct another minimum of three face-to-face contact attempts.

In the sequential mixed-mode design, all sample units were sent an invitation to participate via web first. 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. In case of 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 CAPI. They were contacted at least three times by a face-to-face interviewer on different days and different time periods.

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

Fieldwork for both SM and MM surveys started simultaneously and was conducted by the same fieldwork agency. The fieldwork periods were also equally long. For a more detailed overview of both survey designs please see Chapter 4. Table 5.1 shows the number of completed interviews and response rate (AAPOR RR_1) per survey condition, separately for each ethnic group.

Table 5.1
Number of completed interviews (n) and response rate (AAPOR RR_1) for Single-mode (SM) and Mixed-mode (MM), separately for each ethnic group

	Turkish		Moroccans		Surinamese		Antilleans	
	n	RR_1 (%)	n	RR_1 (%)	n	RR_1 (%)	n	RR_1 (%)
SM	815	52.1	829	48.0	780	41.0	863	44.2
MM Total	533	54.5	556	51.7	515	43.1	537	44.4
MM_Web	210	21.5	245	22.8	250	20.9	260	21.5
MM_Cati	38	3.9	13	1.2	71	5.9	54	4.5
MM_Capi	285	29.1	298	27.7	194	16.2	223	18.4

Tailor-made response-enhancing measures

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

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

The second type of tailoring was 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 interviewers with a shared ethnic background in the area.
Extensive measures were taken to reduce the potential of measurement variance due to the use of different modes, cultural differences and different language versions of the survey. This involved the use of a TRAPD (An acronym for Translation, Review, Adjudication, Pre-testing and Documentation) procedure for the Turkish and Moroccan questionnaires (Harkness 2007), Dillman's unimode approach to questionnaire design (Dillman 2000; 2007), cognitive interviews and the use of simple language. For instance, the Dutch questionnaire had been 'translated' into simpler Dutch (B1-level) by a specialized company. The B1-level corresponds to a language proficiency level that can be readily understood by 95% of the Dutch speaking population. For each of the four ethnic groups, the outcomes of ten survey questions will be analyzed (see Table 5.2 for an overview). These questions vary in content, question type and cognitive effort. The content varies from sociocultural and ethnicity-integration attitudes and statements to more structural measures such as education, home ownership and being a member of the labour force. The question type varies between factual, attitudinal or evaluative questions. The level of cognitive effort varies from answering about one's current situation, stating the degree to which one agrees with a statement or counting the number of times in a fixed period the respondent exhibited a certain behaviour. Depending on the type of variable, we analyze the proportions, mean and variance for the existence of measurement and selection effects.

Table 5.2
Overview of variables included in the analysis

Variable	Question content	No. of answering categories	Scale type	Type of Question	Type of cognitive effort
Man responsible	The man should decide on money issues.	5 + no answer	Ordered categorical	Attitudinal	Degree
Women stop work	A woman should stop working when she has had a baby.	5 + no answer	Ordered categorical	Attitudinal	Degree
Attend religious service	How often do you attend a religious service?	4 + no answer	Ordered categorical	Factual	Count
Interethnic contact	How often do you interact with Dutch in a social context?	3 + no answer	Ordered categorical	Factual	Degree
Language difficult	How often do you experience difficulties with the Dutch language when you have to speak Dutch?	4 + no answer	Ordered categorical	Evaluative	Degree
Opportunity Ethnic	In the Netherlands you get all the opportunities you need.	5 + no answer	Ordered categorical	Evaluative	Degree
Self-identification	With which group do you identify yourself more? Dutch or <ethnic group>?	5 + no answer	Ordered categorical	Evaluative	Degree
Education level	Maximum attained education level.	7	Ordered categorical	Factual	Current
Home owner	Is the house you live in a rented house or a private property?	2	Categorical	Factual	Current
Labour force	Is the respondent a member of the labour force?	3	Categorical	Factual	Current

5.3.2 Method

Several methods have been proposed to deal with inference in sequential mixed-mode surveys, such as the calibration of mode proportions to fixed proportions in the case of repeatedly conducted, sequential mixed-mode, cross-sectional surveys (Buelens and van den Brakel 2011), multiple imputation to disentangle mode and selection effects (Suzer-Gurtekin et al. 2012; Kolinekov and Kennedy 2014), assessing mode effects through re-interviewing and different regression approaches (Jackle et al. 2010), or latent class analysis (Biemer 2001). We will use the method proposed by Vannieuwenhuyze et al.

(2010; 2012) for disentangling mode and selection effects in order to assess measurement effects in a sequential mixed-mode survey, because we also have data available from a single-mode reference survey.

The Vannieuwenhuyze-method requires a single-mode survey to be conducted in parallel with a mixed-mode survey among the same target population. The method has two assumptions with respect to the *respondent* sample (i.e., the responding units in the sample) of both surveys. The first assumption is called the *representativity assumption* and postulates equal coverage error and nonresponse bias in survey estimates coming from both samples. It assumes that no differences in the coverage of the target population are introduced by the sample design, sample frame or the different survey designs. Furthermore, it is also assumed that the potential nonresponse bias on substantive variables is equal in size and direction between the surveys. In other words, the sets of respondents are comparable (Vannieuwenhuyze et al. 2012).

A direct check of the assumption about equal nonresponse bias is seldom available, since nonresponse bias is item specific and not survey specific and the answers of the nonrespondents on substantive variables are per definition unknown (Groves and Peytcheva 2008). A common approach for dealing with nonresponse is weight adjustment with respect to several auxiliary variables assumed to be correlated with the substantive variables (Särndal and Lundström 2010). Vannieuwenhuyze (2010; 2012) suggests methods for detecting indications of potential nonresponse bias, such as the comparison of response rates or a comparison of the *respondent* samples with respect to several auxiliary variables and/or mode-insensitive variables. An efficient method would be to compare the composition of *respondent* samples by using representativity indicators (Schouten et al. 2009).

The second assumption concerns equal measurement error and bias for the reference mode A for both the mixed-mode and the single-mode sample. For instance, interviewer induced measurement errors are equal in size and direction within the face-to-face mode between both surveys.

Vanniewenhuyze's-method is based on the *rule of total probability* or *rule of elimination* from probability theory. The method uses this rule to show that both the mode and the selection effect on a substantive variable can be estimated based on the information available in both surveys. The selection effect of reference mode A on the mean of a substantive variable y [Sa(μ_y)] is defined as the difference between the mean measured by the same mode, but observed on the two different groups of respondents, namely those who would answer by mode A ($\mu_{y_{a|a}}$) and those who would answer by mode B ($\mu_{y_{a|b}}$) in the mixed-mode sample (Vannieuwenhuyze et al. 2010, p. 1030). The part ($\mu_{y_{a|b}}$) is obviously unknown, but based on the rule of total probability and the representativity assumption, it is estimable with the help of the mean on variable y in the single-mode A survey (μ_{ya}) and the proportion of respondents that choose mode A in the mixed-mode survey (τ_a). [equation 1].

$$\mathrm{Sa}(\mu_y) = \mu_{ya|a} - \mu_{ya|b} = \frac{1}{1-\tau_a} * (\mu_{ya|a} - \mu_{ya}) \qquad [1]$$

The mode-effect of mode B on the mean of a substantive variable y Mb(μ_y) is defined as the difference between the estimates obtained by the two different modes, though observed on the same group of respondents, so the mean of mode B respondents in mode A ($\mu_{yb|a}$) and the mean of mode B respondents in mode B($\mu_{yb|b}$). Here ($\mu_{yb|a}$) is unknown, but estimable with the help of the mean on variable y in the single-mode A survey (μ_{ya}) [equation 2].

$$\text{Mb}(\mu_y) = \mu_{yb|a} - \mu_{yb|b} = \frac{1}{1-\tau_a} * \mu_{ya} - \frac{\tau_a}{1-\tau_a} * \mu_{ya|a} - \mu_{yb|b}. \qquad [2]$$

A similar approach is taken to estimate the mode and selection effects on variances and proportions. A detailed description on the estimation of the mode and selection effects on variances and proportions, as well as variance approximation needed for inferences falls outside the scope of this article and for this we refer to Vannieuwenhuyze et al. (2010; 2012) and Vannieuwenhuyze and Molenberghs (2010).

A limitation of this method is the fact that it has been developed in order to disentangle two modes only. Vannieuwenhuyze et al. (2010) also analysed a single-mode CAPI survey and a WEB-CATI-CAPI sequential mixed-mode survey. In that study they chose to combine the WEB and CATI modes and to compare them with the CAPI mode. We feel that combining the CATI and WEB mode may distort the results too much, since the presence or absence of an interviewer is viewed as a major cause of measurement differences (Couper 2011; De Leeuw 1992; 2005; Dillman et al. 2009; Jäckle et al. 2010; Pierzchala 2006; Tourangeau 2000).

Therefore, we prefer to combine CAPI and CATI and compare it to WEB. We are aware that the choice to combine CAPI and CATI violates one of the requirements of the method, namely that one of the modes from the sequential mixed-mode design needs to be identical to the single-mode survey (i.e., CAPI). However, we believe this violation is relatively minor, as the effect of CATI and CAPI interviewers on the measurement of substantive variables is expected to be more similar than the effect of CATI interviewers and WEB. In mode comparison, it has repeatedly been found that there is a dichotomy of modes with and modes without an interviewer (Groves 1989) and meta-analyses show that self-administered forms clearly differ from interviewer administered forms (De Leeuw 1992; Tourangeau et al. 2013). For instance, several studies have shown that social desirability bias is weakest in self-completion modes (such as WEB), stronger in CAPI interviews, and strongest in CATI interviews (Bowling 2005; Kreuter et al. 2007). Furthermore, the number of CATI respondents is very low; depending on the ethnic groups, only two to fourteen percent of the total response (i.e., percentage of the total number of achieved interviews in the mixed-mode samples) is generated by CATI (see Table 5.1). The limited impact on the violation of the measurement equivalence assumption on the overall results by combining CATI with CAPI was also demonstrated when we conducted a WEB-CATI versus CAPI comparison and obtained almost identical results to the ones that we will discuss in this article (Appendix 5.D).

The decision to combine CATI and CAPI also allows us to better estimate the (combined) measurement effect introduced by the combination of modes and TMREM since the TMREM (*bilingual* interviewers with a shared ethnic background and translated questionnaires) were mostly used in CATI and CAPI[3].
Another option would have been to drop all the CATI interviews from the analysis and reweight the samples with respect to several auxiliary variables. However, that would make the required representativity assumption very difficult to meet. In this study, the CATI respondents involve a very selective group of respondents: known fixed landline owners, WEB-nonresponders, first generation immigrants, older immigrants (See Chapter 4).
The existence of relevant measurement and/or selection effects can sometimes not be determined given the sample size (Vannieuwenhuyze 2010; 2012). Similarly to Vannieuwenhuyze (2010; 2012) we do not only assess whether an effect is significant, but we also look at the size of the effect. For interpretation of the results we distinguish between three groups of effect sizes, based on all effects that are detectable given the size of our sample, with a minimal power of 0.8 and a significance level of 0.95 (two-sided). We consider a (positive or negative) effect of less than 5% of the range of the variable (or less than 5% percentage points in case of separate categories) to be a negligible to small effect. An effect of 5% to 10% will be seen as a small to moderate effect, an effect of 10 to 15% a moderate to large effect and an effect is considered large if it is at least 15% of the range of the variable.

5.3.3 Analysis plan and hypotheses

Different causes for measurement effects

The current study includes a sequential mixed-mode design and several TMREM's (i.e., the use of translated questionnaires, bilingual interviewers and interviewers with a shared ethnic background). So, if we expect both assumptions to hold and use Vanniewenhuyze's-method, a detected measurement effect could be the result of a mode effect, a TMREM effect (i.e., an unintended systematic difference introduced by translation, by the use of another language or by being interviewed by someone with a shared ethnic background), a conjunction between mode and one or more TMREM effects, or the result of several TMREM effects together. Furthermore, if we use Vanniewenhuyze's-method and the assumptions do not hold, a detected measurement effect might be the result of a violation of one or both of the method's assumptions or a violation combined with a real measurement effect (i.e., the result of a mode and/or TMREM effects).
With respect to the assumptions, it is unknown how serious a small violation of (one of) the assumption(s) would alter the results or subsequent conclusions. Furthermore, the assumptions are variable by variable assumptions, so a serious violation of the method's assumption for one substantive variable among one ethnic group does not invalidate the

[3] A translated Web-questionnaire (in Turkish and Moroccan-Arabic) was available but only used by 44 Turkish respondents and 9 Moroccan respondents.

results of detected measurement effects on other substantive variables, nor the result on the substantive variable among the other groups. It should, however, make one more cautious when interpreting the results.

To distinguish between the different causes for an observed measurement effect, we test for measurement effects on the combined ethnic groups and re-do the analyses separately for each of the ethnic groups. We chose to first combine and then separately re-do the analysis not for the sake of being able to detect rather small measurement effects for which the combined groups might provide us with the required sample size, but mainly to detect different patterns of measurement effect across the various ethnic groups. With respect to the patterns across the various ethnic groups we expect to find distinct patterns related to the underlying cause (i.e., mode effect, TMREM, etc.) of the measurement effect. This allows us to assess whether or not the detected measurement effect was in fact a mode effect, a translation/language effect, a shared ethnic background effect, a violation of one of the assumptions or a combination thereof. With respect to the underlying causes of the observed measurement effect we expect the following patterns to emerge:

Mode effect patterns

We expect a mode effect to occur systematically across ethnic groups and to have a more or less similar effect size in each of the separate groups. Therefore, we will only conclude that a detected measurement effect on the combined groups is actually a mode effect if the combined estimate and the four separate group estimates are of similar size.

Translation and/or language effect

We expect a translation and/or language effect to occur among the Turkish or the Moroccan sample only. Furthermore, we expect the measurement effect to occur only among these samples, since the observed effect would be language/translation specific and there would be no measurement effect in the combined group analysis. We can compare the translation with the source question and, to make sure this is not a violation of the representativity assumption, we can compare the samples using an external data source, if available, such as registry data. At the same time, we expect to find translation and/or language related measurement effects more often among the Moroccan sample. First of all because of the use of a phonetically translated questionnaire of a spoken only language, which is difficult to read out by interviewers. Secondly, because in translating the questionnaire also a simpler version of classic Arabic was used (Moroccan-Arabic). This language is more commonly used in daily conversation and is more easily understood by non-Dutch speaking Moroccan respondents. The downside of this choice is the lack of nuance in Moroccan-Arabic, so when it comes to ordered answering categories, it may inadvertently lead to measurement differences.

Effect of a shared ethnic background

In case of an effect due to being interviewed by someone with a shared ethnic background, we expect a significant measurement effect to occur only on ethnicity and integration-related questions (i.e., questions on interethnic contact, opportunities

for ethnic minorities, and ethnic self-identification). Furthermore, we expect those measurement effects to be significant and of similar size among the Turkish and the Moroccans samples, but small and non-significant among the Surinamese and Antilleans samples, because of the difference in frequency of use of interviewers with a shared ethnic background (only 50% for the Surinamese and Antillean samples versus 100% for the Turkish and Moroccan samples).
In particular, we expect measurement effects as a result of ethnic background to occur among the Turkish and Moroccans for questions concerning difficulty with the Dutch language, attending religious services, and questions regarding gender roles. For instance, research has shown that language problems are mostly found among Turkish and Moroccans (see Chapter 3). Also, almost all Turkish and Moroccans consider themselves to be religious and more often hold traditional views with respect to the role of men and women (Huijnk and Dagevos 2012). We expect being religious and having traditional values to be viewed as socially desirable within these groups and, as a result, to be overreported in the case of interviews conducted by interviewers with a shared ethnic background.

A combined mode and translation/language effect

We expect that, if the underlying cause of the observed measurement effect is a combination of a mode effect and a translation/language effect, it should show in a distinct pattern when analysing the combined and separate estimates. At least the two ethnic groups that were interviewed in Dutch (the Surinamese and the Antilleans) and one of the other ethnic groups should show similarly sized measurement effects. This is because translation/language effects should be group specific.

A combination of mode effect and effect of shared ethnic background

In case of measurement effects resulting from a combined mode and shared ethnic background effect, we expect a significant measurement effect to occur for the total sample and, furthermore, to show a pattern of significant and similarly sized effects among the Surinamese and Antillean samples and a significant and larger, but similarly sized effect among the Turkish and Moroccan samples. Furthermore, we only expect this to occur on the ethnicity related questions and on questions with a higher likelihood for culturally expected social desirability bias among Turkish and Moroccans.

A combination of translation/language effect and effect of shared ethnic background

In this scenario, several different measurement effect patterns can occur across the groups, that depend both on the direction of the translation/language effect and on the shared background effect. For instance, one may expect a pattern that shows a significantly different measurement effect to occur between the Turkish and the Moroccans and a more or less similar, but smaller measurement effect among Surinamese and Antilleans. The direction of the measurement effect for interviewers with a shared ethnic background should be the same for Surinamese and Antilleans and for either the Turkish or the Moroccan group, without the translation/language effect.

Violations of the representativity or equal measurement error assumption

When significant measurement effects are not detected in the combined analyses, but only in one of the separate analyses, it is likely an indication of a violation of the representativity assumption. This is especially the case with Antilleans and Surinamese, where a translation/language effect is not probable. In case of a mode effect combined with a serious violation of the representativity assumption for one of the ethnic groups, one would expect a more or less similar pattern across the three ethnic groups without a violation of the representativity assumption and a different (i.e., an absent, much larger or even reversed) effect in the group with a serious violation of the representativity assumption, depending on the direction of the violation. It would also be possible to have a combination of either mode or TMREM effects with a minor violation of the representativity assumption. However, in that instance one would need an external data source in order to determine the sources of the measurement effect.

As for a violation of the equal measurement error assumption, we expect that for the CAPI interviews the assumption will not be violated: the same group of CAPI interviewers, the same (translated) CAPI questionnaire, the same interviewer training and the same fieldwork period were used. Next, although this is very difficult to verify without an external source, it seems unlikely for this assumption to be seriously violated for the CAPI and CATI combined mode, where again the same questionnaires were used, but mostly because of the limited number of added CATI interviews. Finally, we partly test for violation of the equal measurement error assumption by taking into account two distinct possible causes for increased measurement variability (i.e., differences as a result of ethnic background and translation/language).

Selection effects

In order to distinguish selection effects from potential violations of the assumptions, we expect to find for the former the same pattern as for a mode effect, that is, a more or less equal effect on the combined and across the separate groups. In case of group specific selection effects, we assume the effect to be the result of violations of the assumptions unless the mode profile (i.e., sociodemographic composition of the respondents within a mode), combined with previous research, offers a plausible explanation.

5.4 Results

Respondent profiles for the different survey modes: A necessary aid to the interpretation of potential measurement and selection effects

A common finding in single-mode survey research is that women participate in surveys more often than men (Groves and Couper 1998; Stoop 2005). This is also the case in this mixed-mode experiment (see Korte and Dagevos 2011). However, our main interest is not whether a specific group is overrepresented in the final sample; we primarily want to find out whether there is a mode preference across the respondents that correlates to the sociodemographic variables. For instance, is there a difference in the level of overrepresentation of female respondents between survey modes in this mixed-mode design? Differences in the composition of sociodemographic variables of respondents across the different modes can aid our understanding why mode and selection effects on substantive variables occur.

A logistic regression including all four ethnic *respondent* samples (i.e., respondents) was conducted to predict the preferred survey mode within the sequential mixed-mode survey (Table 5.3)[4]. In this analysis, the WEB respondents are compared to the respondents that have participated via CATI or CAPI (CAPI+). The predictors included in the model are: *Ethnicity of the respondent* (4 categories, with Moroccans as the reference category), *Female respondent* (dummy), *1st generation immigrant respondent* (dummy), (the natural log of) *Age of respondent* and respondent lives in a *Large city* (dummy).

Table 5.3
logistic regression results on the sequential mixed-mode survey preference of the respondent (WEB = 0 and CATI + CAPI = 1) based on ethnicity, gender, immigration generation, (the natural log of)age and municipality size

Predictor	Coefficient (se)
Ethnicity (reference category = Moroccan)	
Turkish	0.157 (0.126)
Surinamese	-0.238 (0.125)
Antillean	-0.263 (0.127)*
Female	-0.149 (0.089)
1st generation immigrant	0.418 (0.117)**
Ln_Age	0.464 (0.130)**
Large city	0.033 (0.094)
Intercept	-1.555 (0.425)**
Model Fit	
LR chi 2	81.63
df	7
Prob. > chi2	0.0000
Pseudo R2	0.0277
N	2,141

Note * p=<0.05; ** p=< 0.01

The analysis shows that in comparison to the WEB respondents, the CAPI+ respondents are more likely to be first generation immigrants and older, but there is no significant effect for *Gender* and *Large city*. Regarding ethinicity, only Antillean respondents are more likely to participate via WEB than CAPI+ compared to the Moroccan respondents.

[4] See Appendix 5.A for the results of a multinomial logistic regression comparing WEB, CATI and CAPI respondents.

Is the representativity assumption met?

With respect to the representativity assumption, the survey designs do not cause differences in coverage error among the same population. The sampling design and the sampling frame are identical and the use of CAPI with (*bilingual)* interviewers with a shared ethnic background in both survey designs ensures that even the sampled persons that are most difficult to survey have an equal opportunity to participate. On the other hand, the Representativity-indicator (see Schouten et al. 2009) used to check the representativity of the *respondent* samples of both survey designs shows significant differences among all four ethnic groups (see Chapter 4). It turns out that the *unweighted* single-mode *respondent* sample from each ethnic group is significantly more representative with respect to gender, age, municipality size and first or second immigration generation. Also, the single-mode *respondent* samples shows lower estimated maximum nonresponse bias than their sequential mixed-mode counterparts. At the same time, the response rate tends to be significantly higher in the mixed-mode samples, with the exception of the Antilleans (See Chapter 4).

In order to meet the representativity assumption for both *respondent* samples in each ethnic group, a raking procedure (Kalton and Flores-Cervants 2003) has been used to correct for the observed differences in representativity on sociodemographic variables. The *respondent* samples have been weighted to the population distributions on *Gender, Single or multi household, Municipality size, 1st or second generation immigrants* and twelve age groups. As a result, both weighted datasets are comparable on these sociodemographic characteristics.

However, the representativity assumption is a variable by variable assumption, and therefore one can never be completely sure the assumption holds for *all* variables in the survey after the weight adjustment. Nonetheless, the variables used in the weight adjustment are all known to relate to the substantive variables included in the analysis (see for example, Huijnk and Dagevos 2012 or Huijnk et al. 2014) and, consequently, the weight adjustment should increase the representativity and comparability of the *respondent* samples.

As a first check on whether or not the representativity assumption is supported after the weight adjustment, several variables believed to be mode-insensitive are compared between the weighted *respondent* samples. The variables *Age of the partner* and *Household size* are compared within each ethnic group and neither shows any significant differences between the weighted *respondent* samples (appendix 5.B).

Another way to check the tenability of the representativity assumption is by using an external data source, such as registry data. We had limited access to the Dutch registry data. Therefore, as a second check of the representativity assumption, we were able to link the status of each sampled person from both surveys on *Home ownership, Income* (quintiles) and *Socioeconomic category* [working, receiving benefits (social, unemployment, etc.), pension, etc.]. Subsequently, we were able to compare the weighted estimates of *Home ownership, Income* and *Socioeconomic category* between *respondent* samples (Table 5.4).

Table 5.4
Check on the representativity assumption for the weighted *respondent* samples (i.e., responding units) of both survey designs (single-mode and mixed-mode) using *Home ownership*, *Income* and *Social-economic category* from the Dutch registry information, separately for each ethnic group

	Home ownership			Socioeconomic category			Income (quintiles)			n	
Ethnicity	X^2	df	p	X^2	df	p	X^2	df	p	SM	MM
Turkish	6.539	2	0.038	3.570	4	0.467	2.286	5	0.808	815	533
Moroccans	4.506	2	0.105	5.241	4	0.263	8.810	5	0.117	829	556
Surinamese	3.144	2	0.208	8.801	4	0.066	4.956	5	0.421	780	515
Antilleans	2.532	2	0.208	9.620	4	0.047	12.770	5	0.026	863	537

For the Turkish group, there is a significant difference (p = <0.05) between the two *respondent* samples with respect to *Home ownership*, but the other two variables show no significant differences. For the Moroccan and Surinamese *respondent* samples, all estimates show no significant differences between the survey designs, which supports the representativity assumption of the weighted data. As mentioned before, the representativity assumption is a variable by variable assumption, and therefore one can never be completely sure it will hold for all variables in the survey. In the case of the Antilleans, the first check on the representativity assumption (i.e., using mode-insensitive variables) supports the assumption, but the second check (i.e., using registry data) shows that the representativity assumption is not met for two out of three variables. The results of the mixed-mode analysis on the substantive variables among Antilleans should therefore be interpreted with caution and judged on their plausibility.

Measurement and selection effects on substantive outcomes

Table 5.5 presents a summary of the results on measurement and selection effects. The complete results are provided in Appendix 5.C. The first column refers to the variable and the type of question that was analysed. The second column shows where a measurement effect was found: the mean (μ), variance (σ^2), no answer category/not applicable (N.A.) and/or on at least one other category (p_i). Columns three to seven show the size and direction of the measurement effect for the combined groups (C), the Turkish group (T), the Moroccan group (M), the Surinamese group (S) and the Antilleans (A). Column eight shows the main reason for the measurement effect (M= mode effect; E = shared ethnic background effect; V= violation of the representativity assumption and T = translation/language effect). The last five columns show the size and direction of the selection effects. The size and direction of the measurement and selection effect are indicated as follows: '-' or '+' for a (negative or positive) negligible to small effect; '- - ' or '+ +' for a small to moderate effect; '- - -' or '+ + +' for a moderate to large effect and '- - - -' or '+ + + +' to indicate a large effect (see section 5.3.2 for definitions of effect size). Furthermore, an asterisk (*) indicates whether or not the size of the sample is large enough to detect a measurement effect with a 0.95 significance level and a power of 0.8.

On the variable *Man responsible* (see Table 5.2) no consistent pattern of measurement and/or selection effects is detected, there is only a large selection effect among Moroccans. This would mean that Moroccan CAPI+ respondents hold a more traditional view on who should decide on money issues than the WEB respondents. This is in itself is not unlikely, considering the survey mode profile of the CAPI+ respondents, but the same effect is not found among any of the other ethnic groups.
The variable *Women stop work* shows a significant and small to moderate effect on the mean and variance for the combined group. The pattern of small to moderate effects on the mean and variance is also quite consistent across the ethnic groups, which suggests an actual mode effect is behind the measurement effect. The use of face-to-face (and telephone) interviewers causes respondents on average to report more traditional opinions on the role of women with children. It seems likely the effect does not always reach significance because of insufficient sample size rather than it not being a mode effect. No significant selection effect on the mean and/or variance is detected for the combined group analysis, however they are observed on the variance for some of the groups.
The combined group results on *Attend religious service* show small to moderate measurement effects on the separate categories and a moderate to large effect on the average frequency of attending a religious service. The pattern is quite similar across ethnic groups – thereby presenting rather convincing evidence of mode effects – although the large effect among Moroccans also suggests the ethnic background of the interviewer is partly responsible for CAPI+ respondents reporting more frequently that they attend a religious service every day. No selection effects on the combined groups are detected, only effects in opposite direction for the Surinamese and Antilleans. This leads to the conclusion that no systematic selection effects with respect to *Attend religious service* are present across groups.

Of the four ethnicity and integration related questions (*Interethnic contact*, *Language difficult*, *Opportunity ethnic* and *Self-identification*) included in the analysis, *Opportunity ethnic* shows no detectable mode and/or selection effect and is therefore not included in Table 5.5. *Interethnic contact* does not show a significant measurement effect for the combined analysis and only one isolated measurement effect for the Moroccan group, which suggests it is either a translation/language effect or a violation of the representativity assumption. Furthermore, *Interethnic contact* shows a small to moderate, but significant selection effect for the combined group analysis, which is also consistent across the separate groups. This means that CAPI+ respondents interact less with native Dutch in their spare time than the web respondents. Small to moderate measurement effects are found in the combined analysis for both *Language difficult* and *Self-identification*. The consistent pattern across the different ethnic groups suggests that the effect on *Language difficult* is caused by a mode effect, whereas the pattern across the different ethnic groups for *Self-identification* suggests the effect is more likely the result of the shared ethnic background with interviewers.

Table 5.5
Overview of measurement and selection effects on the mean (μ), variance (σ^2), no answer/not applicable category (n.a.) or one or more categories (p_i) found on each of the substantive variables, for the combined ethnic groups and separately for each ethnic group

		Measurement effects						Selection effects				
Variable							Main					
name	Effect on:	C	T	M	S	A	Cause	C	T	M	S	A
Man responsible	μ									----*		
Women stop work	μ	--*	--	---*	--	--	M					
	σ^2	++*	+	++	+	++*	M	-		---*		---*
Attend religious Service	μ	---*	--	----*	----*	-	E+M				+++*	---*
	$p_{>every\ day}$	++*	++	+++*	+		E					
	$p_{>once\ a\ month}$	++*	++	+	++	++*	M					
	$p_{>once\ a\ year}$	--*	--	--	--	-	M					
	n.a.									-*		
Interethnic contact	μ			++*			T or V	++*	++	+	+	+
	σ^2								+*			
Language difficult	μ	--*	-	--*	-	--	M					
	$p_{\text{do not speak Dutch}}$	+*	++	+	+	+*	M					
	σ^2				---*		M					
Self-identification	μ	--*	---*	---*	-	--	E+M					
	σ^2	-	---*	-	---*	++	M					
Education level	μ							---*	----*	--	---*	---*
	n.a.							---*	--- *	---*	--	--
Labour force	p_{employed}	--	-	----*	++	---*	T+V				++++*	
	$p_{\text{not in lf}}$	++	+	+++*	-	+++*	T+V					
Home owner	p_{refusal}									--*		
	p_{rent}	+++*	++++*	+	+	+++	T+V	++	---*	+++*	+++*	++
	p_{owner}	---*	---*	---*	-	--	T+M+V	--	+	---*	---*	--

Note. * significant on 0.95 significance level (2-sided) with a power of 0.80. '–' or '+' = A (positive or negative) negligible to small effect; '--' or '++' = a small to moderate effect; '---' or '+++' = a moderate to large effect; '----' or '++++' = a large effect. M = mode effect; E= shared ethnic background with interviewer; V= violation of the representativity assumption and T= translation/language effect.

On the more structural variables measuring *Education level*, being a member of the *Labour force* and *Home ownership*, we find quite consistent measurement and selection effects among the combined and separate groups. No significant measurement effects are detected on *Education level*, but there is a rather consistent moderate to large selection effect present on the combined and separate groups. CAPI+ respondents have a lower educational level than the WEB respondents.

Only a small to moderate and non-significant measurement effect is found on *Labour force* in the combined group analysis, despite it being highly significant among Moroccans and Antilleans. It is non-significant mainly because of a reversed effect among the Surinamese. However, looking at the different pattern of the Surinamese with respect to *Labour force* for both the measurement and the selection effect leads us to believe that in this instance the representativity assumption is violated and that this is, in fact, a mode effect. The presence of an interviewer causes CAPI+ respondents to report less often that they are employed and to report more often that they are not part of the labour force. This effect is quite possibly the result of the interviewer instructions, in which it was clearly stated that if respondents only work and/or assist in the family business and are not paid as employees they are not considered as part of the labour force.

The question on *Home ownership* shows measurement and selection effects. With the exception of the Turkish group, the selection effect for the combined and separate groups has quite a consistent pattern. If we take into account the violation of the representativity assumption with respect to *Home ownership* among the Turkish samples (see Table 5.4), we may conclude that for the other three groups CAPI+ respondents are more often tenants than owners.

A moderate to large measurement effect is also detected for the combined groups, implying that home owners more often claim that they are tenants when asked by an interviewer. Furthermore, the effect is also present to some degree across the other ethnic groups, albeit not always significant, which can be the result of insufficient sample sizes. This makes it difficult to dismiss the observed measurement effect found on the combined groups as mainly the result of a large violation the representativity assumption among the Turkish. This is particularly the case given that Moroccans also show a significant and moderate to large measurement effect, while for this group the representativity assumption is not violated (see Table 5.4).

Home ownership is one of the variables we were able to link to Dutch registry data. We were able to compare the answers on ownership status of the Turkish and Moroccan respondents with the actual ownership status in 2009 and check whether the measurement effect that we found only masks a violation of the representativity assumption, whether it is truly a measurement effect (caused by mode, translation, etc.) or a combination of the two. Table 5.6 shows the answers on the home ownership question in both single-mode and mixed-mode unweighted *respondent* samples for the Turkish and the Moroccan groups compared to the actual status according to the Dutch registry data in 2009[5]. The columns show the reported status in 2010-2011 and the rows show the actual status in 2009. The frequencies in bold italic show the number of mismatches between the reported status in 2011 and the actual status in 2009.

[5] Excluding missing answers, refusals to answer and other.

Table 5.6
Home ownership status according to the Single-mode and Mixed-mode data for the Turkish and Moroccan groups compared to the actual status according to the Dutch registry data in 2009.

Ownership status in 2011 according to:

	SM						MM					
Ownership status according to the Dutch Registry in 2009	Turkish	No	Yes	N	OR	95%-CI	Turkish	No	Yes	N	OR	95%-CI
	No	483	***39***	522			No	278	***26***	304		
	Yes	***67***	211	278	3.93	3.51 - 4.36	Yes	***39***	174	213	2.40	1.87 – 2.93
	Total	550	250	800			Total	317	200	517		
	Moroccan	No	Yes	N	OR	95%-CI	Moroccan	No	Yes	N	OR	95%-CI
	No	661	***15***	676			No	426	***14***	440		
	Yes	***33***	94	127	15.47	14.82 - 16.12	Yes	***17***	92	109	5.62	4.88 – 6.37
	Total	694	109	803			Total	443	106	549		

There is always a possibility that the situation has actually changed in the meantime. For instance, it is estimated that 5.7% of the Turkish population moved between 2009 and 2011. However, most movers retain the same ownership status after the move. Furthermore, there is no reason to assume any large difference in the frequency of movers between the two survey designs, since the sampling date, sampling frame and reference date are identical.

As also indicated by the results presented in Table 5.5, Turkish and Moroccan home owners seem to report far more often that they are tenants than tenants reporting the opposite (Table 5.6). The odds of a Moroccan home owner in 2009 reporting that they are a tenant are 15.47 times (=661*33/ 15*94) higher than they are for a tenant to report being a home owner according to the answers of the Single-mode (SM) respondents. The effect is smaller, but also significant among Turkish home owners.

The odds of misreporting ownership status among SM Turkish and Moroccans home owner respondents are also significantly higher compared to the same odds in the MM. This indicates that misreporting is also partly measurement related. Additional analysis showed that the measurement effects are mainly the result of translation and interpretation differences, and the lack of instructions in case a respondent did not know the answer in the Web version of the questionnaire.

5.5 Conclusion and Discussion

In this study we investigated the impact of different modes and tailor-made response-enhancing measures (TMREM) on the measurement of ten substantive variables in surveys among non-Western minorities in the Netherlands. In order to detect possible measurement effects introduced by the different modes and/or TMREM, we used a recently developed technique for disentangling mode and selection effects (Vannieuwenhuyze et al. 2010; Vannieuwenhuyze et al. 2012).

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 (WEB-CATI-CAPI) and one sample to single-mode CAPI design; the single-mode was used as reference sample. This resulted in eight groups for analysis. Both survey designs involved the use of TMREM such as the use of translated questionnaires, bilingual interviewers and interviewers with a shared ethnic background. Fieldwork was conducted simultaneously and by the same fieldwork agency.

During the analysis stage we first applied the Vannieuwenhuyze-method on the combined samples of the ethnic groups. Secondly, we conducted the same analysis separately for each ethnic group. Together, the combined and separate results allowed us to distinguish between different underlying reasons for the observed measurement effect. More specifically, we focused on the 'overall' effect of the combined group and the different patterns of the effect across ethnic groups.

The analysis of the combined and separate groups revealed rather consistent measurement effects for seven out of ten variables. The observed patterns tell us these measurement effects are the result of mode, TMREM and violations of the assumptions underlying Vanniewenhuyze's-method which will be discussed in more detail below. Furthermore, these effects were found despite the fact extensive measures were undertaken to minimize mode effects and translation effects. It is likely for mixed-mode surveys that are less carefully designed to suffer from more and larger effects. Sometimes significant measurement effects were not detected in separate groups, but, as Vannieuwenhuyze et al. (2010; 2012) already noted, this also has to do with sample size. Sometimes the sample size is not large enough to detect significant measurement effects of a certain effect size. Also, in one instance – *Labour force* – no significant measurement effect was found for the combined analysis. Most likely this was caused by a serious violation of the representativity assumption for one of the ethnic groups, thereby 'watering down' the actual measurement effect. However, the results for the other three ethnic groups suggest a small to moderate measurement effect also affected this estimate.

With respect to question content, we can draw the following conclusions about the underlying causes for the observed measurement effects. First of all, the respondent's answers on two out of three sociocultural oriented questions are affected by mode. In these instances the web version does seem to elicit more 'honest' responses.

Secondly, only two out of the four integration-ethnicity related questions – *Language difficulty* and *Self-identification* – shows significant measurement effects for the combined groups. The observed pattern across the groups indicates the effect on *Language difficulty* is a mode effect and the effect on *Self-identification* is the result of the shared ethnic background of the interviewer. In case of the latter, the shared ethnic background causes a larger number of respondents to report identifying more with their own ethnic group, which confirms previous results of the effects caused by the ethnicity of the interviewer (see for instance, Schaeffer 1980).

Thirdly, two out of three of the more structural questions – *Labour force* and *Home ownership* – showed rather consistent measurement effects which were mainly the result of violations of both assumptions. For both variables the representativity assumption was violated in the case of one ethnic group. Furthermore, the equal measurement error assumption was possibly violated for *Home ownership* as a result of translation differences. In case of *Labour force,* this effect was caused by specific interviewer instructions. However, one cannot completely dismiss the existence of an actual mode effect for *Labour force.* Home-owners seem more likely to claim they are tenants in the presence of an interviewer.

Our hypothesis regarding the number of measurement effects among Moroccans is confirmed. The most significant measurement effects were found among Moroccans, which suggests the use of a phonetically translated questionnaire and the use of a simpler version of (written) Arabic increase the likelihood of measurement differences. Some anecdotal evidence is available that interviewers find it very difficult to read out phonetic transcriptions.

The analysis of the combined and separate groups also reveals the presence of a rather systematic selection effect on three of the variables (*Interethnic contact, Educational level* and *Home ownership*). Also in this instance the overall effect on one variable – *Home ownership* – is somewhat diluted as a result of a violation of the representativity assumption of Vanniewenhuyze's-method in one group. However, the selection effects are mostly in the expected direction. For instance CAPI and CATI-respondents (CAPI+) tended to be lower educated. This is plausible, since older and 1st generation immigrants are more likely to have responded via CAPI+. The education level among the older respondents and 1st generation immigrants is generally lower compared to the 2nd generation immigrants and younger respondents (Gijsberts et al. 2012). Also, it is plausible for CAPI+ respondents to spend less free time with the native Dutch. The WEB respondents are more often younger and second generation immigrants, which makes them more likely to speak Dutch, since they attended school in the Netherlands. Both factors increase the likelihood of befriending native Dutch.

Insofar as our underlying question is concerned – is the combination of a sequential mixed-mode survey and TMREM among non-Western minorities in the Netherlands a good alternative to a single-mode face-to-face survey with TMREM in terms of measurement error or variability? – the answer is: yes and no. What makes it hard to answer this question is the fact that the use of bilingual interviewers and the use of interviewers with a shared ethnic background are confounded with mode. From the perspective of reducing the social desirability, the web version does seem to elicit more 'honest' responses. This is in accordance with previous research (Heerwegh 2009). Therefore a case can be made for the use of a sequential mixed-mode survey, for example WEB and CAPI.

However, it is evident from our results that the use of different modes introduces systematic measurement differences. Furthermore, our results show that the TMREM also contribute significantly to these differences. The use of interviewers with *a shared ethnic background* seems to elicit more answers that can be viewed as socially desirable within the ethnic group compared to interviewers without a shared background.

Therefore, the use of multiple modes *in combination with* TMREM does increase measurement variability compared to the single-mode with TMREM.

The use of *bilingual* interviewers remains necessary among populations with language barriers. Among populations without any significant language barriers, the benefits of using interviewers with a shared ethnic background are more difficult to assess. If we assume that the use of an interviewer with shared ethnic background leads to more socially desirable answers than the use of an interviewer without the same ethnic background, that might still not undermine the purpose of the survey. In our case, the surveys were meant to assess the comparative socioeconomic status and cultural integration of different ethnic groups. The comparability of the different ethnic groups might actually be reduced if the interviewers with shared ethnic background were only used in certain ethnic (sub)groups.

Harkness (2007) showed the necessity of protocols for careful translation. Our results confirm this and emphasize the importance for situations in which the language is a spoken-only language. When a phonetic transcription is used, additional training should be provided to the interviewers. Another possibility reducing the effect of a translation into a non-written language is the use of AUDIO-CAPI with voice recording of the questions. However, this is only possible if there are not to many regional incomparable dialects

Regarding the usability of Vanniewenhuyze's-method, we were able to provide insight into the tenability of the underlying assumptions and the method's ability to detect measurement and selection effects. As is the case with nonresponse bias, the representativity assumption is variable dependent and not survey dependent. As a result, a violation of the representativity assumption in one instance does not make the method invalid for outcomes of other questions in the same survey.

Our analysis showed that even in the case the recommended strategies, such as nonresponse adjustment and difference testing on mode insensitive variables, it could very well happen that the representativity assumption was not met with respect to other variables. A comparison of three different survey variables with registry data showed that the representativity assumption was violated two out of three times among Antilleans and once for the Turkish group. A further analysis of the Turkish (and Moroccan) answers on *Home ownership* revealed that not only the representativity assumption was violated, but most likely also the equal measurement error assumption, as a result of translation and interpretation differences. Also, a mode effect occurred due to instructions not present in Web.

The 'pattern' approach facilitated the identification of spurious measurement and selection effects as a violation of assumptions. However, one can never be sure if an actual mode or selection effect is partly clouded as a result of a violation. For instance, it is interesting to have found no selection effects on the variable measuring language difficulty given the survey mode profile of the CAPI+ respondents.

A limitation of the current study is the choice of combining CATI and CAPI responses in one group. We are aware it increases the likelihood of a violation of the equal measurement error assumption. However, we believe that, in this instance, a violation of the equal measurement error would more likely lead to an overestimate or more pronounced measurement effects due to mode and TMREM, yet it would less likely to lead to an improbable or hidden effect. Furthermore, the WEB + CATI versus CAPI comparison analysis suggests that the CATI impact is not only small because of the actual number of interviews, but also more in line with the effect of CAPI instead of WEB.

Simulation could provide further insight into the degree to which a violation of the representativity assumption affects the measurement effects. For example, in order to determine the impact on the measurement effect, small to large infractions on the assumption can be simulated also using different sample sizes. All in all, Vanniewenhuyze's-method can be useful in determining measurement differences introduced by mode or other survey design decisions, but it should be applied with caution.

Appendices

Appendix 5.A. Survey mode profile results of the multinomial logistic regression analysis

The following results show the respondent profiles for the different modes within the mixed-mode survey. A multinomial logistic regression including all four ethnic *response* samples was conducted to predict the (relative) preferred survey mode within the sequential mixed-mode survey (Table A.1). In this analysis, the WEB respondents were used as the base category to compare CAPI and CATI. The auxiliary variables included in the model are: *Ethnicity of the respondent* (4 categories with Moroccans as the reference category), *Female respondent* (dummy), *1st generation immigrant respondent* (dummy), (the natural log of) *Age of respondent* and respondent living in a *Large city* (dummy).

The CATI-WEB comparison shows us that Turkish, Surinamese and Antillean respondents are more likely than Moroccan respondents to participate via CATI compared to WEB. Furthermore, the odds of the CATI respondent being female instead of male is higher compared to WEB, just like the odds of the CATI respondent being a first generation immigrant respondent. Finally, the probability of the CATI respondent being older rather than young compared to WEB increases with the (natural log of) age of the respondent.

The CAPI-WEB comparison shows us that Surinamese and Antillean respondents are less likely than Moroccan respondents to participate via CAPI compared to WEB. Also, the odds of being male instead of female are higher in CAPI than in WEB. The same is true for first generation immigant and older respondents.

In terms of survey mode profiles, this means that Surinamese and Antillean respondents are more likely to participate via WEB and CATI than Moroccan respondents. Turkish respondents are also more likely to participate via CATI than Moroccan respondents, but not more likely to participate via WEB. Furthermore, CATI respondents are more likely to be female, older and first generation immigrant respondents compared to WEB respondents. CAPI respondents are more likely to be male, older and first generation immigrant compared to WEB respondents. As a result, this makes the WEB respondent more likely to be second generation immigrant and younger compared to either CATI or CAPI respondents. Also, the WEB respondent is more likely to be male than female compared to a CATI respondent and more likely to be female than male compared to a CAPI respondent.

Table A.1
Multinomial logistic regression results on the sequential mixed-mode survey mode preference of the respondent (WEB = base), based on *Ethnicity*, *Gender*, *Immigration generation*, (the natural log of) *Age* and living in a *Large city*

Comparison	Predictor	Coefficient (se)
CATI vs. WEB	Ethnicity (reference category = Moroccan)	
	Turkish	1.145 (0.338)**
	Surinamese	1.251 (0.326)**
	Antillean	1.494 (0.320)**
	Female	0.352 (0.174)*
	1st generation immigrant	0.569 (0.243)*
	Ln_Age	1.089 (0.251)**
	Large city	-0.039 (0.180)
	Intercept	-7.342 (0.893)**
CAPI vs. WEB	Ethnicity (reference category = Moroccan)	
	Turkish	0.083 (0.127)
	Surinamese	-0.397 (0.129)**
	Antillean	-0.504 (0.133)**
	Female	-0.235 (0.092)*
	1st immigration generation	0.402 (0.122)**
	Ln_Age	0.355 (0.135)**
	Large city	0.049 (0.097)
	Intercept	-1.173 (0.442)**
Model Fit	LR chi 2	180.50
	df	14
	Prob. > chi2	<0.000
	Pseudo R2	0.0458
	N	2,141

Note * p=<0.05; ** p=< 0.01

Appendix 5.B. Results of the comparison between SM and MM samples of the weighted survey estimates *age of the partner* and *household size*, separately for each ethnic group.

Table B.1
Results of mean difference test of the weight adjusted survey estimate *Average household size* between the SM and MM sample, separately for each ethnic group

Ethnicity	SM (s.e)	MM (s.e)	P
Turkish	3.430 (0.065)	3.593 (0.098)	0.1663
Moroccans	3.699 (0.067)	3.601 (0.082)	0.3568
Surinamese	2.748 (0.052)	2.635 (0.061)	0.1561
Antilleans	2.658 (0.055)	2.616 (0.077)	0.6610

Table B.2
Results of mean difference test of the weight adjusted survey estimate *Mean age of partner* between the SM and MM sample, separately for each ethnic group

Ethnicity	SM (s.e)	MM (s.e)	P
Turkish	41.14 (0.511)	41.55 (0.640)	0.6117
Moroccans	43.90 (0.597)	43.29 (0.790)	0.5391
Surinamese	45.07 (0.667)	46.40 (0.876)	0.2262
Antilleans	40.93 (0.741)	40.46 (0.860)	0.6795

Appendix 5.C. Results on the mode and selection-effects for CAPI+ versus Web, for the combined ethnic groups and separately for each ethnic group

Table C.1
Results on mode (M) and selection (S) effects on the mean ($\hat{\mu}$), variance ($\hat{\sigma}^2$) for question “the man should decide on money issues”, combined and separately for each ethnic group

	Combined	Turkish	Moroccans	Surinamese	Antilleans
Man Responsible	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)
M($\hat{\mu}$)	-0.128 (0.072)	-0.290 (0.164)	0.086 (0.150)	-0.037 (0.128)	-0.305 (0.125)
S($\hat{\mu}$)	-0.153 (0.087)	0.260 (0.194)	-0.705 (0.176)	-0.128 (0.161)	0.124 (0.154)
M($\hat{\sigma}^2$)	0.037 (0.103)	0.111 (0.201)	-0.150 (0.247)	-0.034 (0.217)	0.081 (0.172)
S($\hat{\sigma}^2$)	-0.023 (0.122)	-0.123 (0.221)	0.008 (0.254)	0.103 (0.269)	-0.131 (0.221)

Table C.2
Results on mode (M) and selection (S) effects on the mean ($\hat{\mu}$), variance ($\hat{\sigma}^2$) for question “A woman should stop working when she has had a baby”, combined and separately for each ethnic group

Woman stop work	Combined Effect (s.e.)	Turkish Effect (s.e.)	Moroccans Effect (s.e.)	Surinamese Effect (s.e.)	Antilleans Effect (s.e.)
M($\hat{\mu}$)	-0.347 (0.069)	-0.328 (0.155)	-0.529 (0.141)	-0.306 (0.133)	-0.290 (0.122)
S($\hat{\mu}$)	-0.018 (0.085)	-0.245 (0.188)	0.055 (0.162)	0.139 (0.168)	0.169 (0.136)
M($\hat{\sigma}^2$)	0.291 (0.098)	0.164 (0.218)	0.284 (0.168)	0.198 (0.220)	0.402 (0.164)
S($\hat{\sigma}^2$)	-0.182 (0.118)	0.038 (0.250)	-0.515 (0.187)	-0.038 (0.283)	-0.459 (0.214)

Table C.3
Results on mode (M) and selection (S) effects on the proportions and corrected mean, combined and separately for each ethnic group

Attend religious Service	Combined Effect (s.e.)	Turkish Effect (s.e.)	Moroccans Effect (s.e.)	Surinamese Effect (s.e.)	Antilleans Effect (s.e.)
M($\hat{p}_1$) (n.a.)	-0.037 (0.025)	-0.068 (0.031)	-0.025 (0.020)	-0.019 (0.056)	-0.058 (0.055)
M($\hat{p}_2$)(every day)	0.052 (0.015)	0.062 (0.035)	0.151 (0.045)	0.021 (0.013)	-0.007 (0.007)
M($\hat{p}_3$)(at least once a week)	0.036 (0.026)	0.005 (0.064)	0.053 (0.057)	0.087 (0.041)	0.003 (0.044)
M($\hat{p}_4$)(at least once a month)	0.070 (0.018)	0.083 (0.043)	0.038 (0.040)	0.064 (0.028)	0.094 (0.031)
M($\hat{p}_5$)(at least once a year)	-0.077 (0.026)	-0.082 (0.059)	-0.123 (0.055)	-0.072 (0.049)	-0.029 (0.042)
M($\hat{p}_6$)(never/less than once a year)	-0.044 (0.025)	-0.000 (0.056)	-0.090 (0.052)	-0.081 (0.050)	0.007 (0.047)
M($\hat{\mu}\_c$)	-0.418 (0.064)	-0.250 (0.129)	-0.710 (0.120)	-0.599 (0.133)	-0.098 (0.129)
S($\hat{p}_1$) (n.a.)	-0.049 (0.028)	-0.023 (0.028)	-0.048 (0.012)	-0.043 (0.069)	0.021 (0.069)
S($\hat{p}_2$)(every day)	-0.008 (0.019)	0.022 (0.048)	-0.095 (0.056)	-0.013 (0.017)	0.018 (0.013)
S($\hat{p}_3$)(at least once a week)	0.034 (0.033)	0.050 (0.077)	0.036 (0.071)	-0.076 (0.051)	0.083 (0.059)
S($\hat{p}_4$)(at least once a month)	-0.038 (0.022)	-0.110 (0.048)	0.047 (0.053)	-0.037 (0.037)	-0.065 (0.040)
S($\hat{p}_5$)(at least once a year)	0.072 (0.031)	0.079 (0.070)	0.079 (0.066)	0.069 (0.062)	0.044 (0.053)
S($\hat{p}_6$)(never/less than once a year)	-0.012 (0.030)	-0.018 (0.065)	-0.019 (0.060)	0.100 (0.064)	-0.111 (0.055)
S($\hat{\mu}\_c$)	0.021 (0.077)	-0.060 (0.152)	0.207 (0.145)	0.523 (0.163)	-0.424 (0.159)

Table C.4

Results on mode (M) and selection (S) effects on the mean ($\hat{\mu}$) and variance ($\hat{\sigma}^2$) for frequency of interethnic contact, combined and separately for each ethnic group

	Combined	Turkish	Moroccans	Surinamese	Antilleans
Interethnic contact	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)
M($\hat{\mu}$)	0.072 (0.044)	0.059 (0.096)	0.359 (0.070)	0.047 (0.079)	0.021 (0.080)
S($\hat{\mu}$)	0.197 (0.054)	0.213 (0.116)	0.194 (0.080)	0.100 (0.100)	0.117 (0.102)
M($\hat{\sigma}^2$)	0.005 (0.031)	-0.079 (0.045)	-0.075 (0.045)	0.059 (0.064)	0.086 (0.070)
S($\hat{\sigma}^2$)	0.058 (0.038)	0.141 (0.051)	0.051 (0.049)	-0.016 (0.080)	-0.003 (0.089)

Table C.5

Results on mode (M) and selection (S) effects on the proportions and corrected mean, combined and separately for each ethnic group

	Combined	Turkish	Moroccans	Surinamese	Antilleans
Language difficulties	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)
M($\hat{p}_1$) (do not speak Dutch)	0.025 (0.009)	0.054 (0.025)	0.011 (0.023)	0.018 (0.090)	0.023 (0.008)
M($\hat{p}_2$)(yes, often)	0.034 (0.017)	-0.024 (0.055)	0.096 (0.042)	0.012 (0.007)	0.051 (0.014)
M($\hat{p}_3$)(yes, sometimes)	-0.007 (0.023)	-0.024 (0.059)	0.125 (0.051)	-0.030 (0.026)	-0.008 (0.042)
M($\hat{p}_4$)(no, never)	-0.052 (0.028)	-0.008 (0.069)	-0.144 (0.062)	-0.000 (0.028)	-0.066 (0.044)
M($\hat{\mu}\_c$)	-0.136 (0.041)	-0.096 (0.108)	-0.264 (0.094)	-0.048 (0.041)	-0.163 (0.064)
S($\hat{p}_1$) (do not speak Dutch)	-0.013 (0.011)	-0.060 (0.028)	0.025 (0.032)	-0.012 (0.012)	-0.020 (0.010)
S($\hat{p}_2$) (yes, often)	0.051 (0.024)	0.218 (0.071)	0.004 (0.055)	-0.015 (0.006)	-0.044 (0.018)
S($\hat{p}_3$) (yes, sometimes)	-0.019 (0.028)	-0.080 (0.067)	-0.038 (0.064)	-0.006 (0.029)	0.009 (0.053)
S($\hat{p}_4$) (no, never)	-0.018 (0.035)	-0.078 (0.082)	0.009 (0.076)	0.033 (0.032)	0.054 (0.056)
S($\hat{\mu}\_c$)	-0.040 (0.052)	-0.173 (0.130)	-0.047 (0.118)	0.073 (0.046)	0.140 (0.080)

Table C.6

Results on mode (M) and selection (S) effects on the mean ($\hat{\mu}$) and variance ($\hat{\sigma}^2$), combined and separately for each ethnic group

	Combined	Turkish	Moroccans	Surinamese	Antilleans
Opportunity	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)
M($\hat{\mu}$)	-0.134 (0.069)	-0.245 (0.153)	-0.185 (0.136)	-0.060 (0.142)	-0.050 (0.126)
S($\hat{\mu}$)	-0.084 (0.087)	-0.146 (0.182)	0.040 (0.160)	0.000 (0.189)	-0.268 (0.163)
M($\hat{\sigma}^2$)	-0.130 (0.078)	-0.009 (0.181)	-0.055 (0.145)	-0.507 (0.153)	0.061 (0.142)
S($\hat{\sigma}^2$)	0.272 (0.096)	0.079 (0.218)	-0.110 (0.150)	1.107 (0.233)	0.167 (0.178)

Table C.7
Results on mode (M) and selection (S) effects on the mean ($\hat{\mu}$) and variance ($\hat{\sigma}^2$), combined and separately for each ethnic group

	Combined	Turkish	Moroccans	Surinamese	Antilleans
Self-identification	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)
M($\hat{\mu}$)	-0.268 (0.068)	-0.399 (0.124)	-0.393 (0.111)	-0.100 (0.123)	-0.266 (0.133)
S($\hat{\mu}$)	-0.323 (0.080)	-0.090 (0.137)	-0.360 (0.126)	-0.457 (0.150)	-0.064 (0.158)
M($\hat{\sigma}^2$)	-0.074 (0.091)	-0.462 (0.152)	-0.029 (0.116)	-0.544 (0.168)	0.233 (0.168)
S($\hat{\sigma}^2$)	-0.174 (0.104)	0.015 (0.159)	-0.195 (0.111)	0.388 (0.192)	-0.537 (0.213)

Table C.8
Results on mode (M) and selection (S) effects on the mean ($\hat{\mu}$), variance ($\hat{\sigma}^2$) and not applicable, because persons still attending education (n.a.) category for highest achieved educational level, combined and separately for each ethnic group

	Combined	Turkish	Moroccans	Surinamese	Antilleans
EducLevel	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)
M($\hat{\mu}$)	-0.080 (0.153)	0.404 (0.329)	-0.630 (0.330)	0.397 (0.255)	-0.625 (0.254)
S($\hat{\mu}$)	-1.036 (0.180)	-1.564 (0.368)	-0.409 (0.389)	-0.945 (0.299)	-0.820 (0.308)
M($\hat{\sigma}^2$)	-0.049 (0.508)	-0.623 (1.759)	0.281 (0.764)	-0.384 (0.715)	0.489 (0.711)
S($\hat{\sigma}^2$)	-0.346 (0.539)	-0.176 (1.815)	-0.421 (0.857)	-0.393 (0.761)	-0.501 (0.764)
M(*n.a.*)	0.023 (0.025)	0.036 (0.053)	-0.050 (0.052)	-0.035 (0.043)	0.007 (0.049)
S(*n.a.*)	-0.103 (0.029)	-0.129 (0.058)	-0.121 (0.060)	-0.093 (0.050)	-0.062 (0.060)

Table C.9
Results on mode (M) and selection (S) effects on the proportions, combined and separately for each ethnic group

	Combined	Turkish	Moroccans	Surinamese	Antilleans
Labour force	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)
M($\hat{p}_1$) (employed)	-0.064 (0.031)	-0.027 (0.069)	-0.196 (0.064)	0.113 (0.059)	-0.145 (0.057)
M($\hat{p}_2$)(unemployed)	-0.012 (0.017)	0.007 (0.032)	0.017 (0.035)	-0.062 (0.035)	-0.008 (0.035)
M($\hat{p}_3$)(not part of labour force)	0.075 (0.030)	0.021 (0.069)	0.179 (0.063)	-0.051 (0.056)	0.163 (0.053)
S($\hat{p}_1$) (employed)	-0.057 (0.038)	-0.113 (0.081)	0.100 (0.078)	-0.289 (0.075)	0.095 (0.071)
S($\hat{p}_2$) (unemployed)	0.005 (0.021)	-0.051 (0.034)	-0.020 (0.042)	0.121 (0.048)	-0.019 (0.042)
S($\hat{p}_3$) (not part of labour force)	0.051 (0.038)	0.163 (0.082)	-0.080 (0.078)	0.167 (0.072)	-0.096 (0.068)

Table C.10
Results on mode (M) and selection (S) effects on the proportions, combined and separately for each ethnic group

	Turkish	Turkish	Moroccans	Surinamese	Antilleans
Home owner	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)
M($\hat{p}_1$) (refusal)	0.003 (0.006)	-0.022 (0.018)	0.048 (0.015)	-0.000 (0.005)	-0.012 (0.010)
M($\hat{p}_2$)(rent)	0.112 (0.029)	0.224 (0.066)	0.062 (0.048)	0.062 (0.059)	0.116 (0.053)
M($\hat{p}_3$)(owner)	-0.104 (0.029)	-0.167 (0.065)	-0.120 (0.046)	-0.054 (0.059)	-0.091 (0.053)
M($\hat{p}_4$)(other)	-0.011 (0.005)	-0.035 (0.016)	0.010 (0.005)	-0.008 (0.008)	-0.012 (0.008)
S($\hat{p}_1$) (refusal)	-0.014 (0.006)	0.001 (0.018)	-0.062 (0.013)	0.003 (0.008)	-0.001 (0.009)
S($\hat{p}_2$)(rent)	0.089 (0.034)	-0.177 (0.078)	0.179 (0.048)	0.163 (0.072)	0.094 (0.063)
S($\hat{p}_3$)(owner)	-0.088 (0.033)	0.066 (0.076)	-0.107 (0.046)	-0.177 (0.072)	-0.096 (0.063)
S($\hat{p}_4$)(other)	0.013 (0.006)	0.050 (0.022)	-0.010 (0.005)	0.010 (0.048)	0.003 (0.006)

Appendix 5.D. Results on mode and selection-effects for CAPI versus Web and CATI (Web+), separately for each ethnic group.

Table D.1
Results on mode (M) and selection (S) effects on the mean ($\hat{\mu}$), variance ($\hat{\sigma}^2$) for question "the man should decide on money issues", separately for each ethnic group

	Turkish	Moroccans	Surinamese	Antilleans
Man Responsible	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)
M($\hat{\mu}$)	-0.251 (0.142)	0.082 (0.143)	-0.029 (0.100)	-0.247 (0.101)
S($\hat{\mu}$)	0.228 (0.176)	-0.712 (0.168)	-0.218 (0.147)	0.067 (0.135)
M($\hat{\sigma}^2$)	0.103 (0.189)	-0.143 (0.203)	-0.026 (0.170)	0.070 (0.142)
S($\hat{\sigma}^2$)	-0.053 (0.235)	-0.028 (0.236)	0.283 (0.241)	-0.186 (0.195)

Table D.2
Results on mode (M) and selection (S) effects on the mean ($\hat{\mu}$), variance ($\hat{\sigma}^2$) for question "A woman should stop working when she has had a baby", separately for each ethnic group

	Turkish	Moroccans	Surinamese	Antilleans
Woman stop work	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)
M($\hat{\mu}$)	-0.289 (0.135)	-0.503 (0.134)	-0.237 (0.103)	-0.237 (0.091)
S($\hat{\mu}$)	-0.239 (0.172)	0.043 (0.156)	0.104 (0.150)	0.070 (0.121)
M($\hat{\sigma}^2$)	0.119 (0.185)	0.263 (0.160)	0.153 (0.171)	0.345 (0.139)
S($\hat{\sigma}^2$)	0.142 (0.223)	-0.520 (0.181)	0.069 (0.260)	-0.378 (0.203)

Table D.3
Results on mode (M) and selection (S) effects on the proportions and corrected mean, separately for each ethnic group

Attend religious Service	Turkish Effect (s.e.)	Moroccans Effect (s.e.)	Surinamese Effect (s.e.)	Antilleans Effect (s.e.)
M($\hat{p}_1$) (n.a.)	-0.059 (0.026)	-0.028 (0.020)	-0.015 (0.043)	-0.047 (0.045)
M($\hat{p}_2$)(every day)	0.054 (0.031)	0.144 (0.043)	0.017 (0.010)	-0.006 (0.006)
M($\hat{p}_3$)(at least once a week)	0.004 (0.056)	0.050 (0.054)	0.068 (0.032)	0.002 (0.036)
M($\hat{p}_4$)(at least once a month)	0.072 (0.037)	0.037 (0.038)	0.049 (0.022)	0.077 (0.025)
M($\hat{p}_5$)(at least once a year)	-0.071 (0.051)	-0.117 (0.052)	-0.056 (0.038)	-0.024 (0.034)
M($\hat{p}_6$)(never/less than once a year)	-0.000 (0.048)	-0.086 (0.049)	-0.063 (0.039)	-0.003 (0.038)
M($\hat{\mu}\_c$)	-0.216 (0.111)	-0.674 (0.115)	-0.467 (0.103)	-0.078 (0.100)
S($\hat{p}_1$) (n.a.)	-0.018 (0.025)	-0.046 (0.011)	-0.074 (0.059)	0.053 (0.062)
S($\hat{p}_2$)(every day)	0.018 (0.043)	-0.087 (0.054)	-0.003 (0.017)	0.019 (0.013)
S($\hat{p}_3$)(at least once a week)	0.032 (0.070)	0.003 (0.068)	-0.052 (0.045)	0.041 (0.051)
S($\hat{p}_4$)(at least once a month)	-0.086 (0.044)	0.057 (0.052)	-0.051 (0.029)	-0.055 (0.035)
S($\hat{p}_5$)(at least once a year)	0.074 (0.064)	0.079 (0.064)	0.048 (0.055)	-0.004 (0.045)
S($\hat{p}_6$)(never/less than once a year)	-0.020 (0.059)	-0.007 (0.058)	0.131 (0.058)	-0.054 (0.050)
S($\hat{\mu}\_c$)	0.248 (0.140)	0.477 (0.139)	-0.259 (0.136)	-0.065 (0.165)

Table D.4
Results on mode (M) and selection (S) effects on the mean ($\hat{\mu}$) and variance ($\hat{\sigma}^2$) for frequency of interethnic contact, separately for each ethnic group

Interethnic contact	Turkish Effect (s.e.)	Moroccans Effect (s.e.)	Surinamese Effect (s.e.)	Antilleans Effect (s.e.)
M($\hat{\mu}$)	0.051 (0.083)	0.196 (0.078)	0.036 (0.062)	0.017 (0.065)
S($\hat{\mu}$)	0.199 (0.106)	0.206 (0.093)	0.168 (0.086)	0.058 (0.088)
M($\hat{\sigma}^2$)	-0.070 (0.042)	-0.131 (0.048)	0.046 (0.050)	0.069 (0.052)
S($\hat{\sigma}^2$)	0.142 (0.050)	0.060 (0.055)	-0.031 (0.069)	-0.084 (0.071)

Table D.5
Results on mode (M) and selection (S) effects on the proportions and corrected mean, separately for each ethnic group

	Turkish	Moroccans	Surinamese	Antilleans
Language difficulties	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)
M($\hat{p}_1$) (do not speak Dutch)	0.047 (0.022)	0.010 (0.022)	0.014 (0.007)	0.018 (0.007)
M($\hat{p}_2$)(yes, often)	-0.020 (0.047)	0.094 (0.040)	0.010 (0.006)	0.041 (0.012)
M($\hat{p}_3$)(yes, sometimes)	-0.020 (0.051)	0.033 (0.051)	-0.023 (0.020)	-0.006 (0.034)
M($\hat{p}_4$)(no, never)	-0.007 (0.060)	-0.138 (0.059)	-0.000 (0.022)	-0.054 (0.036)
M($\hat{\mu}\_c$)	-0.082 (0.094)	-0.252 (0.090)	-0.038 (0.031)	-0.133 (0.052)
S($\hat{p}_1$) (do not speak Dutch)	-0.050 (0.025)	0.028 (0.031)	-0.011 (0.010)	-0.020 (0.006)
S($\hat{p}_2$) (yes, often)	0.185 (0.064)	0.006 (0.054)	-0.012 (0.005)	-0.046 (0.012)
S($\hat{p}_3$) (yes, sometimes)	-0.070 (0.060)	-0.060 (0.061)	-0.005 (0.026)	-0.054 (0.043)
S($\hat{p}_4$) (no, never)	-0.064 (0.074)	0.025 (0.074)	0.028 (0.028)	0.120 (0.045)
S($\hat{\mu}\_c$)	-0.148 (0.118)	-0.038 (0.114)	0.062 (0.040)	0.206 (0.064)

Table D.6
Results on mode (M) and selection (S) effects on the mean ($\hat{\mu}$) and variance ($\hat{\sigma}^2$), separately for each ethnic group

	Turkish	Moroccans	Surinamese	Antilleans
Opportunity	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)
M($\hat{\mu}$)	-0.213 (0. 132)	-0.176 (0.130)	-0.047 (0.109)	-0.040 (0.103)
S($\hat{\mu}$)	-0.226 (0.163)	0.018 (0.155)	0.004 (0.168)	-0.275 (0.145)
M($\hat{\sigma}^2$)	0.000 (0.159)	-0.048 (0.138)	-0.393 (0.119)	0.047 (0.116)
S($\hat{\sigma}^2$)	0.018 (0.196)	-0.058 (0.144)	0.905 (0.195)	0.118 (0.163)

Table D.7
Results on mode (M) and selection (S) effects on the mean ($\hat{\mu}$) and variance ($\hat{\sigma}^2$), separately for each ethnic group

	Turkish	Moroccans	Surinamese	Antilleans
Self-identification	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)
M($\hat{\mu}$)	-0.345 (0.108)	-0.374 (0.106)	-0.077 (0.096)	-0.219 (0.110)
S($\hat{\mu}$)	-0.085 (0.124)	-0.362 (0.121)	-0.489 (0.132)	0.119 (0.138)
M($\hat{\sigma}^2$)	-0.423 (0.107)	-0.040 (0.085)	-0.430 (0.097)	0.146 (0.111)
S($\hat{\sigma}^2$)	0.050 (0.148)	-0.212 (0.107)	0.349 (0.159)	-0.590 (0.184)

Table D.8

Results on mode (M) and selection (S) effects on the mean ($\hat{\mu}$), variance ($\hat{\sigma}^2$) and not applicable, because persons still attending education (n.a.) category for highest achieved educational level, separately for each ethnic group

EducLevel	Turkish Effect (s.e.)	Moroccans Effect (s.e.)	Surinamese Effect (s.e.)	Antilleans Effect (s.e.)
M($\hat{\mu}$)	0.342 (0.277)	-0.598 (0.314)	0.298 (0.192)	-0.488 (0.201)
S($\hat{\mu}$)	-1.289 (0.320)	-0.372 (0.374)	-0.767 (0.257)	-0.596 (0.259)
M($\hat{\sigma}^2$)	-0.369 (1.190)	0.210 (0.706)	-0.193 (0.441)	0.121 (0.463)
S($\hat{\sigma}^2$)	-0.521 (1.254)	-0.392 (0.805)	-0.286 (0.512)	-0.848 (0.531)
M(*n.a.*)	0.036 (0.048)	0.048 (0.049)	0.027 (0.034)	0.030 (0.034)
S(*n.a.*)	-0.101 (0.056)	-0.118 (0.058)	-0.056 (0.045)	-0.059 (0.044)

Table D.9

Results on mode (M) and selection (S) effects on the proportions, separately for each ethnic group

Labour force	Turkish Effect (s.e.)	Moroccans Effect (s.e.)	Surinamese Effect (s.e.)	Antilleans Effect (s.e.)
M($\hat{p}_1$) (employed)	-0.024 (0.059)	-0.187 (0.061)	0.088 (0.046)	-0.118 (0.046)
M($\hat{p}_2$)(unemployed)	0.006 (0.028)	0.017 (0.033)	-0.048 (0.027)	-0.006 (0.043)
M($\hat{p}_3$)(not part of labour force)	0.018 (0.060)	0.170 (0.060)	-0.039 (0.043)	0.124 (0.043)
S($\hat{p}_1$) (employed)	-0.085 (0.073)	0.098 (0.075)	-0.228 (0.065)	0.053 (0.063)
S($\hat{p}_2$) (unemployed)	-0.039 (0.031)	-0.025 (0.040)	0.117 (0.044)	-0.023 (0.037)
S($\hat{p}_3$) (not part of labour force)	0.124 (0.074)	-0.072 (0.075)	0.111(0.063)	-0.031 (0.061)

Table D.10
Results on mode (M) and selection (S) effects on the proportions, separately for each ethnic group

	Turkish	Moroccans	Surinamese	Antilleans
Home owner	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)	Effect (s.e.)
M($\hat{p}_1$) (refusal)	-0.019 (0.016)	0.046 (0.015)	-0.000 (0.004)	-0.010 (0.008)
M($\hat{p}_2$)(rent)	0.194 (0.057)	0.059 (0.046)	0.048 (0.046)	0.094 (0.044)
M($\hat{p}_3$)(owner)	-0.145 (0.056)	-0.114 (0.043)	-0.042 (0.046)	-0.074 (0.043)
M($\hat{p}_4$)(other)	-0.030 (0.014)	0.009 (0.005)	-0.006 (0.007)	-0.010 (0.006)
S($\hat{p}_1$) (refusal)	-0.004 (0.015)	-0.059 (0.013)	-0.002 (0.003)	0.001 (0.008)
S($\hat{p}_2$)(rent)	-0.074 (0.070)	0.168 (0.046)	0.177 (0.062)	0.065 (0.056)
S($\hat{p}_3$)(owner)	0.028 (0.068)	-0.100 (0.044)	-0.187 (0.062)	-0.064 (0.056)
S($\hat{p}_4$)(other)	0.050 (0.022)	-0.009 (0.005)	0.013 (0.012)	-0.002 (0.002)

6 The impact of method bias on the cross-cultural comparability in face-to-face surveys among non-Western minorities in the Netherlands

This Chapter describes a study that investigates the impact of several sources of method bias on the cross-cultural comparison of attitudes towards gender roles and family ties among non-Western minority ethnic groups. In particular, it investigates how interviewer effects, the use of an interviewer with a shared ethnic background, interview language, interviewer gender, gender matching, the presence of others during the interview and differences in sociodemographic sample composition of non-Western minority ethnic groups affect the cross-cultural comparison of attitudes towards gender roles and family ties between these groups.

The data used in this study come from a large scale face-to face survey conducted among the four largest non-Western minority ethnic groups in the Netherlands for which Statistics Netherlands drew a random sample of named individuals from each of the four largest non-Western minority populations living in the Netherlands. Furthermore, methods are introduced to estimate the potential impact of method bias on cross-cultural comparisons.

The results show that measurement of both gender roles and family ties constructs are full scalar invariant across the different ethnic groups, but that observed differences in attitudes between ethnic groups especially towards gender roles are influenced by method bias. This in turn leads to biased comparisons between ethnic groups because of differences in the size of the various sources of method bias, the differential impact of the same method bias between ethnic groups and the combination thereof.[1]

6.1 Introduction

In general population surveys, non-Western minorities – or ethnic minorities as they are sometimes referred to – tend to be underrepresented (Feskens 2009; Groves and Couper 1998; Schmeets and Van der Bie 2005). Ethnic minorities are difficult to survey mainly because of cultural differences, language barriers, sociodemographic characteristics, and a high mobility (Feskens et al. 2010; Feskens et al. 2006; Stoop 2005).

To reduce nonresponse due to language barriers or cultural differences among ethnic minorities, it is often necessary to make use of Tailor-Made Response-enhancing Measures (TMREM). Examples of these TMREM are the use of translated questionnaires, bilingual interviewers, and interviewers with a shared ethnic background (See Chapters 3 and 4; Groeneveld and Weijers-Martens, 2003; Kemper, 1998; Martens, 1999). However, these TMREM may increase the measurement variability of survey estimates. For example, interviewers can systematically affect the way respondents answer survey

[1] This chapter has been published as Kappelhof, J.W.S. (2014). The impact of method bias on the cross-cultural comparability in face-to-face surveys among ethnic minorities. Methods, Data, Analyses, 8, 1, 79-118 DOI: 10.12758/mda.2014.00xx

questions, especially with respect to more sensitive questions (Tourangeau and Yan 2007). Furthermore, the ethnicity of the interviewer and the language of the interview can systematically affect the way respondents answer survey questions as well (Van't Land 2000). Needless to say that potential translation errors in case of translated questionnaires are another source of increased measurement variability.

These TMREM can also affect cross-cultural comparability, for example, if there are differences between the ethnic groups in the number or intensity in which these TMREM were used. Comparability issues can also arise in case the TMREM cause systematic differences between ethnic respondents groups in the way they respond to survey questions (i.e., TMREM have a differential impact). A possible reason would be, for instance, differing attitudes between ethnic groups towards what are sensitive topics (Lee 1993). Also, factors that are not (intended as) part of the survey design can complicate or bias comparisons between ethnic groups if the level or presence of these factors varies between these ethnic groups or has a differential effect. For instance, culturally specific or different response strategies between ethnic groups, such as acquiescence (Billiet and Davidov 2008; Cheung and Rensvold 2000), social desirability (Johnson and Van de Vijver 2003) or extreme response styles (Morren et al. 2012a; Morren et al., 011; Morren et al. 2012b), but also other factors such as the presence of others during the interview, interviewer gender or a gender match between a respondent and an interviewer (Veenman 2002), may generate such effects. Veenman (2002) discusses a range of reasons for which the presence of others during the interview can cause respondents to adjust their answers.

Differences in sample composition of the different groups with respect to important background variables can also complicate the interpretation of observed differences between these groups (Van de Vijver 2003; van de Vijver and Leung 1997). This may cause problems, especially if one is interested in attempting to isolate 'true' cultural differences from differences in sociodemographic composition in which the latter may also affect survey estimates of the various ethnic groups. This can be particularly relevant if one tries to assess the effectiveness of a 'one size fits all' policy on ethnic groups that differ substantially from a sociodemographic point of view.

In the present chapter we investigate how these different factors affect the cross-cultural comparison of two sociocultural integration constructs – attitudes towards *Gender roles* and attitudes on *Family ties*- between non-Western ethnic groups living in the Netherlands. Research suggests that questions about sensitive topics may elicit more measurement bias (e.g., social desirability) via interviewer-assisted modes of data collection (Tourangeau and Yan 2007). Sociocultural integration issues, such as *Gender roles* and *Familiy Ties*, among non-Western ethnic groups in the Netherlands are highly relevant for policy makers. However, the questions measuring these sensitive concepts may suffer from a higher degree of social desirability bias, especially when data is collected via face-to-face surveys. The combination of the topics (gender roles, family ties) and the method of data collection (face-to-face) in our data is therefore suitable for the aim of this study.

This chapter sets out to investigate:

1 how interviewer effects influence the cross-cultural comparison of attitudes on *Gender Roles* and *Family ties* between non-Western groups in the Netherlands; more specifically, the following aspects will be studied:
   1.1 how the use of an interviewer with a shared ethnic background affects the cross-cultural comparison of attitudes on *Gender Roles* and *Family ties* between non-Western groups in the Netherlands;
   1.2 how the language of the interview affects the comparison of attitudes on *Gender Roles* and *Family ties* between non-Western groups in the Netherlands;
   1.3 how interviewer gender and gender matching impact the cross-cultural comparison of attitudes on *Gender Roles* and *Family ties* between non-Western groups in the Netherlands;

2 how the presence of others during the interview affects the comparison of attitudes on *Gender Roles* and *Family ties* between non-Western groups in the Netherlands;

3 to what degree the observed differences in attitudes on *Gender Roles* and *Family ties* between non-Western groups can be attributed to differences in sociodemographic composition between non-Western populations in the Netherlands.

The data used in this study come from a large-scale face-to-face survey conducted between November 2010 and June 2011. Statistics Netherlands drew a random sample of named individuals from each of the four largest non-Western minority populations living in the Netherlands. The next section of this chapter provides an overview of the requirements for conducting valid cross-cultural comparisons and the possible sources of bias that can complicate or invalidate these comparisons. This is followed by the description of the data and methods used to answer our research questions and subsequent results, ending with our conclusion and discussion.

6.2 Sources of bias that can invalidate or complicate cross-cultural comparisons in face-to-face surveys

In recent years, several books describing guidelines and best practices for conducting cross-cultural or cross-national comparative surveys have been published as well as guidelines on how to analyse cross-cultural survey data (see, for example Davidov et al. 2011; Harkness et al. 2010; Stoop et al. 2010). This is understandable, since a multitude of errors and biases can complicate or even invalidate cross-cultural or cross-national comparisons of theoretically based concepts (He and Van de Vijver 2012; Poortinga and Van de Vijver 1987; Van de Vijver and Leung 1997; Van de Vijver and Tanzer 2004).
When it comes to cross-cultural comparisons, a number of equivalence requirements need to be met before meaningful cross-cultural or cross-national comparisons of theoretical concepts can be made. First of all, the intended concept needs to be understood and have meaning in the different countries or cultures. This is commonly referred to as conceptual equivalence (Hui and Triandis 1985; Johnson 1998).

Johnson (1998) refers to the other requirements as forms of procedural equivalence. These forms of procedural equivalence have to do with the way the measurement instrument intended to measure the theoretical concept is constructed and they have a hierarchical structure (Vandenberg and Lance 2000). Three types of measurement equivalence are commonly distinguished for the measurement model (van de Vijver and Leung, 1997; van de Vijver and Tanzer, 2004)[2].

First of all there is construct equivalence. Johnson (1998, p. 9.) refers to this as follows "A measure can be identified as having this type of equivalence to the degree that it exhibits a consistent theoretically-derived pattern of relationships with other variables across the cultural groups being examined." In a multi group confirmatory factor analysis approach this relates to configural equivalence (Hox et al. 2010; Vandenberg & Lance 2000).

Secondly, for cross-cultural or cross-national comparison there is the requirement of equal metric units of the measurement instrument used to measure the concept. This is commonly referred to as measurement unit equivalence, metric invariance or weak factorial invariance.

Thirdly, to ensure fairness and equity of cross-cultural or cross-national comparison of concepts, measurement instruments are not only required to use the same metric, they are also required to have the same origin. This type of equivalence is also referred to as full scalar invariance, measurement invariance, strict factorial invariance or scalar equivalence (Meredith 1993; Meredith and Teresi 2006; Vandenberg and Lance 2000; Wicherts 2007).

Bias in cross-cultural or cross-national comparisons

Three sources of bias that can threaten the validity of cross-cultural or cross-national comparisons are commonly distinguished. These are construct bias, item bias and method bias (Kankaras and Moors 2009; Van de Vijver 2003; Van de Vijver 2011; Van de Vijver and Leung 1997; Van de Vijver and Tanzer 2004). Construct bias occurs when the requirement of construct equivalence is not met. This can happen when non-identical constructs are measured across cultures or countries, or when there is only a partial overlap of the construct between the cultures or countries. Construct bias happens at the level of the measurement instrument designed to capture the theoretical concept.

Item bias happens at the individual question level and occurs when translations of questions (or items) lead to differences in question meaning or ambiguity. Item bias can also be the result of cultural specifics which can be viewed as a form of differential item functioning (DIF) (Mellenbergh 1989). DIF is a term that stems from education testing and happens when persons of equal capability or intelligence arrive at different capability or intelligence scores based on the specific wording of an item.

Method bias happens at survey level and can be introduced by a variety of factors which are distinguished in the following three categories: incomparability of samples,

[2] Some distinguish more than three forms of measurement equivalence and make a distinction between strong (no equal residual variances) and strict factorial invariance (equal residual variances).

administration bias, and instrument bias. Incomparability of samples refers to differences in the sample composition with respect to important sociodemographic characteristics of the respondents. Administration bias refers to bias that is introduced as a result of differences in how the questionnaire is administered (e.g., interviewer effects, presence of others during the interview, interviewer characteristics), differences in questionnaire design, differences in mode of administration, etc. Instrument bias refers to bias that is introduced as a result of differences in familiarity with being interviewed, but also differences in cultural specific answer strategies.

Research into different sources of method bias.

Within cross-cultural or cross-national research, method bias has received relatively little attention in comparison with construct and item bias (Van de Vijver 2011). As far as method bias is concerned, differential answering strategies, such as acquiescence and other types of response styles, appear to have received the most attention (see for instance, Baumgartner and Steenkamp 2001; Billiet and Davidov 2008; Billiet and McClendon 2000; Chen et al. 1995; Cheung and Rensvold 2000; He and Van de Vijver 2013; Hui and Triandis 1989; Johnson et al. 2005; Marin et al. 1992; Morren et al. 2011; Morren et al. 2012a; Morren et al. 2012b; Ross and Mirowsky 1984). This is not surprising, since the respondent is always a part of the survey process.

However, many studies concerned with response styles pay relatively little attention to other sources of method bias that can contribute to the observed differences in response styles, despite the fact that these data are often collected via an interviewer-assisted mode of data collection. For example, the SPVA-study – Social-economic Position of Ethnic groups – aimed to measure the socioeconomic position and sociocultural integration conducted among ethnic minorities in the Netherlands. This study was conducted face-to-face and further research on these data has shown the existence of differential response styles (Morren et al. 2012a; Morren et al. 2011). For its data collection through CAPI, the SPVA survey also used translated questionnaires, interviewers with a shared ethnic background, allowed proxy interviews and family member interpreters (Groeneveld and Weijers-Martens 2003). So, the question is to which degree these differential response styles are the result of characteristics of the respondents themselves and to which degree they are affected by different impacts of interview language, the presence of others during the interview, gender of the interviewer, the ethnicity of the interviewers, proxy interviews and family member interpreters.

Usually, a lack of information on interviewer characteristics and interview setting prevents a more detailed analysis of these types of method bias in cross-cultural research. However, this does not mean that these factors do not bias estimates and, as a result, also lead to biased comparisons. There has been extensive research on the existence of interviewer effects and it has been shown that respondents' answers can be affected by interviewer gender, interviewer race and/or differences (or similarities) between interviewer and respondent such as gender match and race (Anderson et al. 1988; Davis 1997; Davis et al. 2010; Finkel et al. 1991; Rhodes 1994; Schuman and Converse 1971; Williams Jr 1964; Veenman 2002; van der Zouwen 2006). Especially the match between the race of the interviewer and that of the respondent plays a role in the answers given on culturally sensitive questions (Campbell 1981; Cotter et al. 1982; Sudman and Bradburn 1974; Schuman and Converse 1971; Van Heelsum 1997; Van't Land 2000).Furthermore, a meta-analysis on sensitive questions in surveys by Tourangeau and Yan (2007) shows that respondents not only adjust their responses to sensitive questions in the presence of interviewers but also in the presence of others, such as family members.

The incomparability of samples can also bias cross-cultural comparisons (He and Van de Vijver 2012; Kankaras and Moors 2009). Several studies have analyzed the impact of different sociodemographic sample composition of the compared cultural groups on the observed cross-cultural differences (Arends-Tóth and Van de Vijver 2008; Fernandez and Marcopulos 2008; Leung et al. 1998). Several procedures on how to deal with the incomparability of samples, also known as observed heterogeneity, have been proposed (Boehnke et al. 2011; Lubke et al. 2003; Lubke and Muthen 2005) as well as other procedures to separate compositional differences from 'true' group differences (DiNardo et al. 1996; Huang et al. 2005; Oaxaca 1973).

6.3 Data and Methods

6.3.1 Data

The data used in this chapter come from the Dutch Survey on the Integration of Minorities (SIM) that sets out to measure the socioeconomic position of non-Western minorities as well as their sociocultural integration. It is a nationwide, cross-sectional, face-to-face CAPI survey; and the fieldwork was conducted by GfK Netherlands between October 2010 and June 2011 among the four largest non-Western minority groups living in the Netherlands plus a Dutch reference group. For this face-to-face survey, Statistics Netherlands drew five samples of named individuals: one random sample was drawn from each of five mutually exclusive population strata; Dutch of Turkish, Moroccan, Surinamese, and Antillean[3] descent and the remainder of the population (mostly native Dutch) living in the Netherlands, aged years and above. The present study focuses on how response-enhancing measures, interview setting, interviewer characteristics and the incomparability of samples in face-to-face surveys can affect cross-cultural comparisons between non-Western ethnic minority groups. This is why the samples containing native Dutch are excluded from this study, the analysis being therefore based on four samples.

The official definition, as is used in statistical research in the Netherlands, of Dutch of Turkish, Moroccan, Surinamese, and Antillean descent includes persons that were either born in Turkey, Morocco, Surinam or the Dutch Antilles[4] or have at least one parent who was born there. In case 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. The four ethnic groups in this

[3] Including Aruba

[4] or Aruba

study make up about two-thirds of the total non-Western population, which amounts to approximately 7% of the total population in the Netherlands (CBS-statline, 2014). For the purpose of brevity, they will be referred to as Turkish, Moroccans, Surinamese and Antilleans in the remainder of this chapter.
The response rate (AAPOR definition 1, (AAPOR, 2011) of the SIM2011 face-to-face survey varied between the four ethnic groups and is shown in Table 6.1. Table 6.1 also includes, the gross sample and the sample size of each of the four *response* samples (i.e., the sample of the respondents).

Table 6.1
Response rate (AAPOR definition 1), *response* sample size and gross sample of SIM2011 face-to-face survey, separately for each ethnic group

Ethnic Group	Response rate (%)	Response sample	Gross sample
Turkish	52.1	815	1,565
Moroccan	48.0	829	1,740
Surinamese	41.0	780	1,930
Antillean (incl. Aruban)	44.2	863	1,974

In this chapter the SIM2011 response data file will be used. The response data file contains respondents' answers to survey questions, but also sociodemographic information on the respondent, sociodemographic information on the interviewer and interviewer observations (Table 6.2). Six survey questions measuring sociocultural integration will be used in this analysis. These questions or a slightly larger set of questions have been used to measure sociocultural integration of non-Western ethnic minorities in the Netherlands for over a decade (Arends-Tóth and Van de Vijver 2008; Dagevos and Gijsberts 2009; Dagevos and Schellingerhout 2003; Dagevos et al. 2007). The first set of three questions aims to measure *Gender role* attitudes and the second set of three questions aims to measure *Family ties*. The interviewer observation data are the result of a short form that an interviewer had to complete after each interview. In this form they had to record in which language the interview was conducted, how well they believed the respondent was able to understand and speak Dutch, but also if there were others present during the interview and if they had, according to the interviewer, influenced the answers of the respondents.

Table 6.2
sim2011 data used in the analysis

Questions on sociocultural integration

- [mangeld] It is best if the man is responsible for the finances. (Ranging from 1= completely agree to 5=completely disagree).
- [inkjongs] It is more important for boys than girls to earn their own money. (Ranging from 1= completely agree to 5=completely disagree).
- [vrwstopw] A woman should stop working when she has child. (Ranging from 1= completely agree to 5=completely disagree).
- [thuishuw] It is best for children to live at home until they get married. (Ranging from 1= completely agree to 5=completely disagree).
- [vertrfama] I trust my family more than my friends. (Ranging from 1= completely agree to 5=completely disagree).
- [kibezoud] Children that live close to their parents' home should visit them at least once a week. (Ranging from 1= completely agree to 5=completely disagree).

Sociodemographic information on the respondent

- Ethnicity (Turkish, Moroccan, Surinamese and Antillean)
- Gender
- Age Group (15-24; 25-34; 35-44; 45-54; 55-64; 64+)
- Immigration generation (first generation immigrant; second generation immigrant)
- Education level (max. primary school; lower secondary; upper secondary; tertiary or more)
- Municipality size (over 250000; between 250000 and 50000; less than 50000)
- Employment status (employed, not employed, not part of the labour force)
- Has children (yes; no)
- Has partner (yes; no)
- Weight variable (design weight plus nonresponse adjustment)

Sociodemographic information on the interviewer

- Unique id number
- Ethnicity of the interviewer (Turkish, Moroccan, Surinamese, Antillean, Dutch)
- Gender of the interviewer

Interviewer observations

- Others present during the interview (no; yes, but no influence; yes, influence)
- In which language was the interview conducted (Dutch; mostly Dutch; half Dutch/half native language; mostly native language; native language)
- What was the respondent's Dutch language proficiency level (good; fair, poor, bad)

Note. Original questions were in Dutch and these are translated by the author.

6.3.2 Hypotheses with respect to the research questions

Interviewer effects

Interviewer dependent correlation between the answers of respondents is not often modeled in cross-cultural or cross-national studies, but it has the potential to affect the cross-cultural comparison when the data is collected face-to-face.

Hypothesis: Observed differences between ethnic groups with respect to *Gender roles* and *Family ties* can be partly explained by interviewer effects.

The effect of bilingual interviewers with a shared ethnic background

Interviewers may have an effect on the responses and especially, the use of bilingual interviewers with a shared ethnic background can impact survey outcomes in several ways. First of all, they can have an effect with respect to potential nonresponse bias. They can interview respondents that would not have participated due to language difficulties in combination with functional illiteracy or cultural etiquettes. Nonresponse bias on survey outcomes would occur if these potential respondents would have a different opinion on those survey topics and they were not able to participate.

Secondly, they can have an effect with respect to potential measurement bias. Here we can distinguish two effects: the interview language and shared ethnic background. Both have the potential to increase measurement bias. For instance, the question delivery or wording of a translated questionnaire can cause a systematic difference which is, of course, intertwined with the translated questionnaire. Also, their shared ethnic background may elicit more responses that are viewed as socially desirable within the ethnic group.

The use of bilingual interviewers with a shared ethnic background in SIM2011 does not allow for this level of disentanglement of bias. For instance, *all* respondents of Moroccan or Turkish origin were interviewed by a bilingual interviewer with a shared ethnic background. This was a necessary step not only because greater cultural familiarity due to a shared ethnic background increases the willingness to respond, but mostly because language difficulties are still quite common among the Turkish and Moroccans. This would allow the respondent to answer either in Dutch or in their native tongue.

About half of the interviews among respondents of Surinamese or Antillean origin were conducted by interviewers with a shared ethnic background, because Dutch is the mother tongue for many, if not all persons of Surinamese or Antillean origin in the Netherlands.

The SIM2011 face-to-face survey data do allow for the estimation of how the use of (bilingual) interviewers with a shared ethnic background affected the cross-cultural comparison with respect to potential nonresponse bias. In the SIM2011 data information was available on the language in which the interview was conducted, the level of the Dutch language skill and the ethnicity of the interviewer (Table 6.2). Here it was assumed that respondents would not have participated because of language problems or cultural differences if the interview was conducted mostly in their native language and the interviewer also assessed that the respondent's Dutch language proficiency level was poor.

A comparison between the model excluding and the one including these respondents will show the impact of the increased nonresponse on the cross-cultural comparison. Hypothesis: The use of bilingual interviewers with a shared ethnic background will have a systematic effect on the cross-cultural comparison. In particular, it will result in more traditional views with respect to *Gender roles* and *Family ties*. First of all, with respect to nonresponse bias we expect respondents who otherwise would not to participate due to language problems or cultural specific reasons to hold more traditional views towards *Gender roles* and *Family ties*. Secondly, we expect that the shared ethnic background elicits more traditional views toward *Gender roles* and *Family ties* because these are felt as more socially desirable within the ethnic group.

The effect of interview language

The SIM2011 data also allows for an estimate of the effect of interview language on the cross-cultural comparison. In this instance, the data about interview language was used to create a dummy indicating whether the interview was conducted (almost) completely in Dutch or not. Not only among Turkish and Moroccans, but also among the Surinamese and Antilleans, some of the interviews were at least partly conducted in another language as well. Obviously, the interview language will be part measurement and part nonresponse related. Furthermore, the effect of the ethnicity of the interviewer will be confounded with the interview language and also potential systematic differences introduced by a translated questionnaire can contribute although that effect should be isolated (i.e., indicator and language dependent).
Hypothesis: Interview language has a systematic effect on the measurement of *Gender roles* and *Family ties*. If the interview language is Dutch, this will lead to less traditional views towards *Gender roles* and *Family ties*.

Interviewer gender and gender match

In the SIM2011 data, information on the interviewer gender as well as the gender of the respondent was available (Table 6.2). This allowed for the construction of both an interviewer gender and a *matched/unmatched* indicator to test how interviewer gender and gender match affect the cross-cultural comparison of sociocultural issues. However, given the topics (*gender roles* and *family ties*) and the traditional views of some of these ethnic groups, we might expect men and women to react differently in the presence of a gender (un)match. For instance, women may give less traditional answers in the presence of a female interviewer whereas men may become more traditional in the presence of a male interviewer. This interaction may be masked if only a *match/unmatched* indicator is fitted. To test this hypothesis an interaction term (gender respondent with gender interviewer) was created in order to find out if there was an effect of interviewer gender and/or differential effect of gender match between men and women.

Hypothesis: Interviewer gender and gender matching will effect the cross-cultural comparability. In particular, we expect that interviews conducted by a male interviewer will result in more traditional views towards *Gender roles* and *Family ties* from the respondents compared to interviews conducted by a female interviewer, especially in the case of male respondents.

The presence (and potential influence) of others

In the SIM2011 data information on the presence of others was available (Table 6.2). This allowed for the construction of a *presence (*dummy) indicator to test how the presence of others affects the cross-cultural comparison of *Gender roles* and *Family ties*. A score of '1' (presence) was assigned to the dummy indicator if the interviewer assessed that a third party present during the interview exerted a direct or indirect influence on the way the respondent answered the questions. In all other instances (i.e., no one present or someone present but no noticeable influence) a score of '0' was assigned to the dummy.

Hypothesis: The presence of others during an interview will systematically affect the results concerning *Gender roles* and *Family ties*.

Incomparability of samples

With respect to the last research question – the incomparability of samples- we expect that part of the observed differences between the ethnic groups can be explained by differences in sociodemographic composition.

6.3.3 Methods

A variety of different modeling and analysis techniques have been used to detect equivalence of measures in cross-cultural research. See Braun and Johnson (2010) for an extensive overview.

In the present study multi group confirmatory factor analysis is used (MGCFA) (Joreskog 1971) to test if the base model – full scalar invariance of the two-factor model of sociocultural integration among the four non-Western minority groups in the Netherlands – adequately describes the data. The latent variable *Gender roles* is measured by the following three items: MANGELD; INKJONGS and VRWSTOPW (Table 6.2). The latent variable *Family Ties* is measured by THUISHUW, VERTRFAMA and KIBEZOUD (Table 6.2).

The full scalar model is used as the basic model (Model 0) and in this chapter we do not focus on the question whether a less restrictive model (e.g., configural equivalence, metric invariance or partially measurement invariant) describes the data better, but rather we focus on the question how method bias can bias the full scalar model with respect to cross-cultural comparisons of sociocultural integration among non-Western minorities in the Netherlands.

The MGCFA analyses have been conducted with Mplus version 6.11 (Muthén and Muthén 2011). Both factors have ordered categorical indicators and therefore the WLSMV (Mean- and Variance-adjusted Weighted Least Square) estimator will be used to address the multivariate normality assumption (Lubke and Muthén 2004).

In addition, several, non-nested models, corresponding to the research questions are going to be analyzed and compared, which normally leads to the use of AIC or BIC fit indices to compare the models (Kuha 2004). However, the combination of WLSMV and the modeling of interviewer effects through clustering does not allow for models to be compared using these indices.[5] Therefore the fit of every model will be judged separately using three often used fit indices: the root mean square error of approximation (RMSEA) (Steiger 1989), the Tucker-Lewis index (TLI) (Tucker and Lewis 1973) and the comparative fit index (CFI) (Bentler 1990).

The root mean square error of approximation (RMSEA) is an absolute fit index that examines closeness of fit. A RMSEA value of more than 0.1 is seen as an indication of poor fit, a value of 0.05 to 0.08 as acceptable and a value below 0.05 as good to very good (Hu and Bentler 1999), although the absoluteness of these cut-off values has been criticized more than once (see for example Chen et al. 2008). The comparative indices “Tucker-Lewis index (TLI)” and “comparative fit index (CFI)” compare the fit of the model under consideration with fit of baseline-model. Fit is considered adequate if the CFI and TLI values are above 0.90, better if they are above 0.95.

Interviewer effects.

This model involves the inclusion of an unique interviewer ID as a cluster variable in the MGCFA test of full scalar equivalence (Model 1). This allows for a correction of possible interviewer-dependent correlation between the answers of respondents that were interviewed by the same interviewer. A comparison between model 0 and model 1 would give an indication as to how possible interviewer effects influence the cross-cultural comparisons of sociocultural integration (i.e., gender roles and family ties) among non-Western minorities in the Netherlands. For the remainder of the analysis, model 1 is chosen to be the reference model, since it more accurately describes the data structure. The interviewer effects will also be included in the remaining models.

Bilingual interviewers with a shared ethnic background: nonresponse

In this instance model 1 will be used, but it will be fitted on a selection of the respondents (Model 2). The respondents that participated in their native language *and* for whom the interviewer assessed that their Dutch language proficiency level was poor were excluded. A comparison between the Model 1 and Model 2 (excluding respondents due to language problems) will show the impact of the increased nonresponse due to language problems on the cross-cultural comparison.

Interview language; the presence of others; interviewer gender and gender match.

Interview language, the presence of others, interviewer gender and gender match are sources of method bias that are not randomly assigned across experimental conditions, but are confounded with respondent’s characteristics. In order to assess if and how these sources of method bias systematically influenced the cross-cultural comparison of *Gender*

[5] Using a maximum likelihood estimator to compare non-nested models based on categorical data would allow the use of BIC. Mplus allows for this approach where instead of a MGCFA, a latent class approach is used with knownclass and type=mixture instead of the grouping variable. However, this does not allow for the modeling of interviewer effects using unique interviewer id as a cluster variable, because that requires type =complex.

roles and *Family ties,* a multiple group MIMIC model (Multiple Indicators Multiple Causes) was used, in which the impact of these sources of method bias, together with eight other sociodemographic variables on the respondent, were regressed on the latent variables and indicators (see Table 6.2: Sociodemographic information on the respondent). This will be referred to as Model 3 (M3) and if there is no systematic bias introduced by these sources of method bias they should not have a significant impact on the latent variables. Furthermore, a comparison between Model 1 en Model 3 will show the impact of these combined types of method bias on the cross-cultural comparison.

The incomparability of samples

The four non-Western groups in this study differ in sociodemographic composition (CBS-statline). A propensity score weighting method is used to investigate how the incomparability of the sociodemographic composition of samples (IoS) between ethnic groups affects cross-cultural comparisons (Bia and Mattei 2008; DiNardo et al. 1996; Huang et al. 2005; Imbens 2000; Rosenbaum and Rubin 1983).
The selection of important sociodemographic variables for the propensity score reweighting was done in three steps. As a first step, ordered logistic regression was used to ascertain which of the eight sociodemographic background variables have a significant effect on the different categorical indicators (see Table 6.2: Sociodemographic information on the respondent). As a second step, a check for significant differences in the composition of the four ethnic groups with respect to these sociodemographic background variables was conducted. As a third step, only those sociodemographic background variables were selected to be included in the propensity score weighting model for which it was shown that they a) have a significant impact on at least one of the categorical indicators and b) show a significant difference between at least two ethnic groups. This led to the propensity score reweighting of the different ethnic groups with respect to four sociodemographic background variables: “Municipality size”, “Employment status”, “Education level” and “Immigration generation”. The comparison of the model with propensity weighted samples (Model 4) with Model 1 would allow for an estimation of the effect of IoS on the observed cultural differences[6].

6.4 Results

Model 0: full scalar invariance

The results of the three fit indices show that full scalar equivalence (M0) has an acceptable fit. This means that both factor means can be compared between the different ethnic groups in a fair and equitable way (Table 6.3)[7].

[6] As a check on the usability of the propensity score weighting method to disentangle ‘true’ cultural differences from IoS on the cross-cultural comparison of sociocultural integration, the Oaxaca-Blinder decomposition (OBD) method was also used (Blinder 1973; DiNardo 2006; Jann 2008; Oaxaca 1973). This should yield similar results (DiNardo 2006).

[7] Response samples are weighted to the respective population distribution for gender, household size, municipality size, immigration generation, age groups (12).

Table 6.3
Fit indices results for each model

Model	RMSEA	$CI^{0.95}_{rmsea}$	CFI	TLI
M0	0.079	0.072 – 0.085	0.940	0.961
M1	0.053	0.047 – 0.060	0.936	0.958
M2	0.055	0.047 – 0.062	0.935	0.958
M3	0.021	0.016 – 0.026	0.938	0.921
M4	0.049	0.043 – 0.056	0.952	0.969

The factor means of *Gender roles* and *Family ties* of the different ethnic groups are shown in Figures 6.1 and 6.2 under M0. Figures 6.1 and 6.2 show the change in relative positions of the factor means of *Gender roles* and respectively *Family ties* among the ethnic groups after correcting for the various sources of method bias. For details on the numerical values of the parameter estimates and their the respective standard errors, see Appendix 6.A. It can be seen that Turkish and Moroccans have, one average, a similar, more traditional attitude towards *Gender roles* and *Family ties* in comparison to the Surinamese and Antilleans, although there is a significant difference in factor mean for *Family Ties* between Turkish and Moroccans (Tables 6.4 and 6.5). There are no significant differences between Turkish and Moroccans for *Gender Roles* as well as no significant differences between Surinamese and Antilleans for both *Gender Roles* and *Family Ties* (Tables 6.4 and 6.5). The remaining group comparisons all show significant differences between ethnic groups for both factor means[8]

Table 6.4
Overview of significant differences between ethnic groups for Gender Roles, separately for each model

Gender roles	M0	M1	M2	M3	M4
T vs. M					
T vs. S	*	*		*	
T vs. A	*	*	*	*	
M vs. S	*	*		*	
M vs. A	*			*	
S vs. A					

Note. * = Bonferroni corrected significance level (0.05/n of tests). T = Turkish, M = Moroccans, S = Surinamese and A = Antilleans

[8] Based on t-test comparison of means for independent groups using a Bonferroni adjusted significant level for multiple comparisons.

Table 6.5
Overview of significant differences between ethnic groups for Family Ties, separately for each model

Family Ties	M0	M1	M2	M3	M4
T vs. M	*	*			
T vs. S	*	*	*	*	*
T vs. A	*	*	*	*	*
M vs. S	*	*	*	*	*
M vs. A	*	*	*	*	*
S vs. A					

Note. * = Bonferroni corrected significance level (0.05/n of tests). T = Turkish, M = Moroccans, S = Surinamese and A = Antilleans

Model 1: The impact of interviewer effects on the cross-cultural comparison

In model 1 (M1), interviewer effects are taken into account when testing for full scalar invariance. The inclusion of interviewer effects where interviewers are modelled as a clustering of observations by unique interviewer number resembles more closely the actual structure of the sample and has a good fit according to the fit indices (Table 6.3). As could be expected, the correction for interviewer effects mainly results in larger standard errors around factor loadings and thresholds for the indicators of both means (See Appendix 6.A). The relative positions of both *Gender Roles* and *Family Ties* of the ethnic groups are only slightly affected, but this does not change the ordering (Figures 6.1 and 6.2). However, there is no significant difference for *Gender Role* anymore between Moroccans and Antilleans (compare M0 and M1 in Table 6.4). This means that the observed difference between Moroccans and Antilleans in Model 0 is the result of interviewer effects.

Figure 6.1
Relative positions on Gender Roles of the ethnic groups

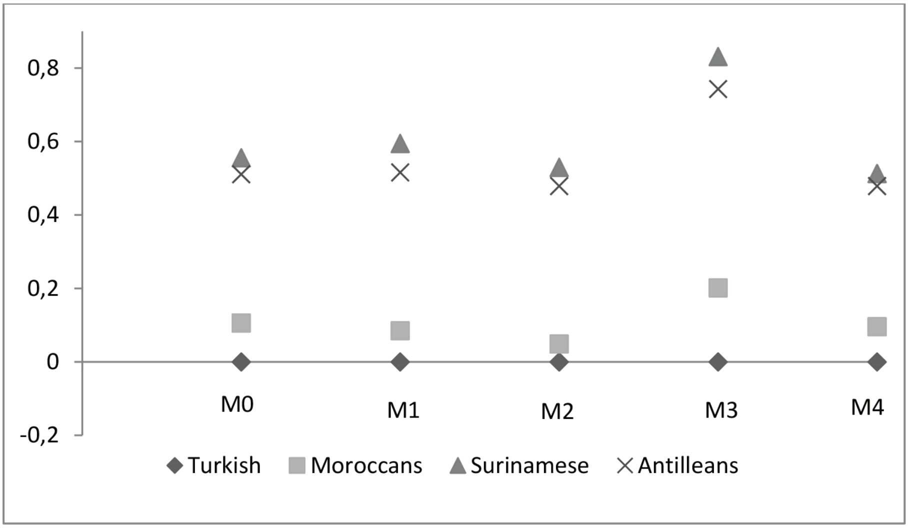
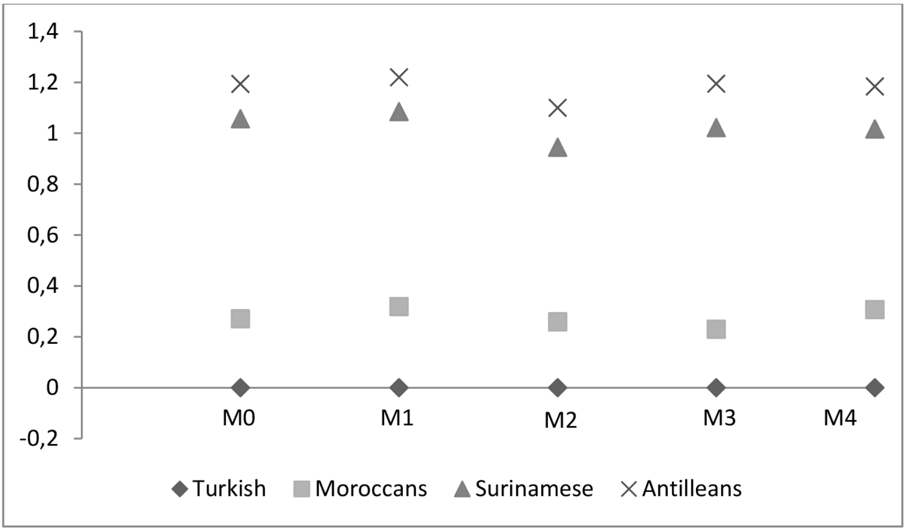


Figure 6.2
Relative positions on Family Ties of the ethnic groups



Model 2: The impact of a (bilingual) interviewer with a shared ethnic background on the cross-cultural comparison in terms of nonresponse bias

The comparison of Model 2 (M2) with Model 1 (M1) shows the impact of a (bilingual) interviewer with a shared ethnic background on the cross-cultural comparison in terms of nonresponse bias. Model 2 also has a good fit according to the fit criteria (Table 6.3). Compared to Model 1, the ethnic groups would have more similar attitudes if no provisions were made to accommodate for persons who do not speak Dutch or have a cultural specific etiquette when it comes to being asked to participate in an interview (see Figures 6.1 and 6.2). For attitudes towards *Gender Roles* only a significant difference between Turkish and Antilleans would remain and for *Family Ties* the observed difference between Turkish and Moroccans would no longer be significant (Tables 6.4 and 6.5). Since the Tailor-Made Response-enhancing Measures (TMREM) mostly affected the Turkish and Moroccans, it can be said that the exclusion of potential respondents due to language problems and lack of cultural etiquette leads to less traditional attitudes of Turkish and Moroccans.

Model 3: The effect of interview language, interviewer gender and gender match interaction, the presence of others on the cross-cultural comparison

Table 6.6 presents the results of the analysis with respect to the impact of *interview language*, *interviewer gender*, *gender match interaction* and *the presence of others* on attitudes towards *Gender roles* and *Family ties*. The complete results can be seen in appendix 6.B. Model 3 (M3) shows an acceptable fit (Table 6.3).

The analysis results show that being interviewed in your native language by a bilingual interviewers with a shared ethnic background significantly affects the attitudes Turkish, Moroccan and Antillean respondents have towards *Family ties*. In all cases more traditional views with respect to *Family ties* are reported. Among the Surinamese there is no significant effect for interview language. This is mostly due to the fact that there are only very few Surinamese interviews conducted in another language.

Table 6.6
The impact of interview language, interviewer gender, gender match and the presence of others on *Gender Roles* (GR) and *Family Ties* (FT), separately for each ethnic group

	Turkish		Moroccans		Surinamese		Antilleans	
	GR	FT	GR	FT	GR	FT	GR	FT
– Interview language		*		*				*
– Interviewer gender			*					
– Gender match	*							
– Others present					*	*		*

Note. * p = <0.05

The *Interviewer gender* only has an effect among Moroccans and only on attitudes towards *Gender roles*. In this instance, Moroccan respondents report less traditional attitudes when the interview is conducted by a female interviewer.

There is an interaction effect for *Gender match* on attitudes towards *Gender roles* among Turkish respondents. Turkish male respondents report more traditional attitudes when the interview is conducted by a male interviewer, while there is no significant effect in the case of Turkish female respondents.

The *presence of others* during the interview significantly affects the attitudes of Surinamese for both *Gender roles* and *Family ties*, as well as Antilleans' attitudes towards *Family Ties*. In all instances the presence of others led to more traditional opinions. Interestingly enough this effect is not (significantly) present among Turkish and Moroccans. The number of interviews in which the interviewer found the presence of others to have a biasing effect varied between 5.6 percent of all interviews conducted among Antilleans and 7.2 percent of all interviews conducted among Surinamese (Turkish 5.8 % and Moroccans 6.4%).

With the exception of attitudes towards *Family ties* among Antilleans, there is at least one significant source of method bias present that systematically affects the attitudes reported by the respondents. Furthermore, there is no source of method bias that has a consistent impact across ethnic groups for one or both latent constructs. As a result, the cross-cultural comparison of these attitudes is biased when comparing the ethnic groups. The actual size of the bias with respect to the cross-cultural comparison of latent means between ethnic groups depends on both the size of the effect and the number of respondents showing this effect.

Model 3 (M3) in figures 6.1 and 6.2 shows the (estimated) relative positions of the latent means for each ethnic group in case adjustments are made for the impact of these sources of method bias. In this case, eight sociodemographic characteristics were also included as covariates to take into account the nonrandom allocation of these source of method bias. Model 3 (M3) in Tables 6.4 and 6.5 show how the adjustments impact the ethnic group comparison. In this instance, the adjustments resulted in the same significant differences as Model 0 (M0) with the exception of the significant difference between Turkish and Moroccans for *Family Ties*.

Model 4: The impact of the incomparability of samples on the cross-cultural comparison

A propensity score weighting method has been used to assess the impact of differences in sociodemographic sample composition between ethnic groups. A summary of the significant differences between the ethnic groups for eight sociodemographic variables is given in Table 6.7 (see Table 6.2 for a description of the sociodemographic variables included in this comparison and Appendix 6.C for the actual results). For modeling reasons, the original variables – *municipality size* and *employment status* – have been condensed to dummies -*Big city dweller* (y/n) and *Employed* (y/n). 21 significant differences are observed between the ethnic groups if they are weighted to their respective population

distributions[9]. Using the propensity weighting procedure described in section 6.3.3, only seven of these significant differences remained, observed on two variables – *Age Group* and *Partner* – that were not included in the propensity score weighting model. The reason for their exclusion from the propensity score weighting model was that these sociodemographic variables did not have a significant impact on the indicators used to measure *Gender Roles* and *Family Ties* (see also Appendix 6.C).
The comparison of Model 4 (M4) with Model 1 (M1) shows the impact of differences in sample composition for five sociodemographic variables (*Immigration generation, Educational level, Big city dweller, Employed* and *Children*, see Table 6.7) between ethnic, non-Western groups on the cross-cultural comparison of attitudes towards *Gender Roles* and *Family Ties*. Model 4 has a good to very good fit according to the criteria (Table 6.3).

Table 6.7
Summary of the significant differences in sociodemographic characteristics between the ethnic groups

Variable (no. of categories)	Weighted to population distribution	Propensity score reweighted
Gender (2)		
Age group (6)	TS*; MS*; SA*	TS*; MS*; SA*
Immigration generation (2)	SA*	
Education level (4)	TS*; TA*; MS*; MA*	
Big city dweller (2)	TM*; TS*; SA*	
Employed (2)	TS*; TA*; MS*; MA*; SA*	
Children (2)	TA*;	
Partner (2)	TS*; TA*; MS*; MA*	TS*; TA*; MS*; MA*

Note: *significant p =<0.01; T = Turkish; M= Moroccans; S=Surinamese and A = Antilleans

The observed differences in attitudes towards *Gender Roles* between the ethnic groups are to some small degree the result of the differences in sample composition; the effect is even less noticeable for *Family Ties,* where differences in sample composition hardly affect the results at all (see Figures 6.1 and 6.2). With respect to *Gender Roles,* the attitudes are more alike when there is a correction for the incomparability of samples, as compared to Model 1, none of the significant differences observed between the ethnic groups persist (Table 6.4). This is not the case for *Family Ties,* where the correction only leads to a non-significant effect between Turkish and Moroccan compared to Model 1 (Table 6.5).

[9] Weighted to the respective population distribution for gender, household size, municipality size, immigration generation, age groups (12)

6.5 Conclusion and discussion

The present study investigated how interviewer effects, the use of an interviewer with a shared ethnic background, interview language, interviewer gender, gender matching, the presence of others during the interview and differences in sociodemographic sample composition of ethnic minority groups can affect the comparison of attitudes towards gender roles and family ties.

The data used in this study comes from a large-scale face-to-face survey conducted between October 2010 and June 2011 for which Statistics Netherlands drew a random sample of named individuals from each of the four largest non-Western minority populations living in the Netherlands. The data contained not only answers to substantive questions, but also sociodemographic information on both respondent and interviewer characteristics, as well as interviewer observations regarding the interview.

As a first step, a multi group confirmatory factor analysis model approach was used to test for full scalar invariance of the two factor model (*Gender roles* and *Family ties*). The model showed an acceptable fit, which meant the latent factor means for both *Gender role* and *Family Ties* could be compared in a meaningful way across the four ethnic groups.

As for the first research question – "How do interviewer effects influence the cross-cultural comparison of attitudes on *Gender Roles* and *Family ties* between non-Western groups in the Netherlands?" – interviewer effects were added to this base model using the unique interviewer number as cluster variable. This reflected the data structure well and the results show that the addition of interviewer effects as cluster variable mostly lead to increased standard errors for all parameter estimates. The effect on the parameter estimates was marginal, which led to some minor changes in the estimated means of *Gender roles* and *Family Ties*. As a result of the increased standard errors and a slight change in the relative position of Moroccans, it was shown that the observed cross-cultural difference on attitudes towards *Family Ties* between Moroccans and Antilleans was mostly the result of interviewer effects. This confirms our hypothesis that the observed differences between ethnic groups with respect to *Gender roles* and *Family ties* can be partly explained by interviewer effects.

The second research question – "How does the use of an interviewer with a shared ethnic background affect the cross-cultural comparison of attitudes on *Gender Roles* and *Family ties* between non-Western groups in the Netherlands?" – was addressed in terms of nonresponse, in which way does the increase in nonresponse due to language problems and cultural differences affect cross-cultural comparison between the ethnic groups? The estimated additional nonresponse as a result of not using bilingual interviewer was based on interview language and the interviewers assessment of the Dutch language proficiency level of the respondent. The analysis showed that the increase in nonresponse had a significant impact on the cross-cultural comparison of *Gender roles.* Without the use of bilingual interviewers with a shared ethnic background, the attitudes towards *Gender roles* turned out to be a lot more similar across the ethnic groups. A specific group of respondents having a more traditional view would have been missed. This means that our hypothesis with respect to the second research question is also confirmed, at least

with respect to nonresponse bias. The use of bilingual interviewers with a shared ethnic background resulted in more traditional views with respect to *Gender roles* and *Family ties.*
The third research question – how does the language of the interview affect the comparison of attitudes on *Gender Roles* and *Family ties* between non-Western groups in the Netherlands- was assessed in combination with other potential sources of method bias. To find out how interview language affected cross-cultural comparison a dummy was made which, together with dummies indicating interviewer gender, gender match, the presence of others as well as eight important sociodemographic variables such as education, gender, age, etc., was regressed as covariate on the latent variables of *Gender roles* and *Family ties*. For this a multi group MIMIC (Multiple Indicators MultIple Causes) model was used. The inclusion of the sociodemographic variables on the respondents was done to correct as much as possible for the inherent confoundedness of these sources of method bias with respondent characteristics.
Interview language had an effect on attitudes towards *Family ties* among Turkish, Moroccans and Antilleans. When interviewed in their native language, they all give (significantly) more traditional opinions. As for Surinamese, no significant effect of interview language was found for either factor. This is not surprising, since only a handful of respondents completed the interview in another language. Also in this instance the hypothesis is confirmed. Interview language has a systematic effect on the measurement of *Gender roles* and *Family ties* and being interviewed in Dutch leads to less traditional views towards *Gender roles* and *Family ties.*
There are several remarks that need to be made in order to place this result of interview language in the right context. First of all, the effect of interview language is confounded with the effect of interviewer ethnicity. However, all Turkish and Moroccan respondents were interviewed by bilingual interviewers with a shared ethnic background, therefore no further disentanglement was possible. On the other hand, some of the interviewer ethnicity effect might already be captured by the modeling of interviewer effects.
Secondly, this effect might also partially be the result of systematic differences introduced by translation. However, the latter is unlikely, since the effect was not detected for just one ethnic group, but for three, one of which never benefitted from a translated questionnaire at all. In addition, the effect was measured on the factor, not on the indicators.
Thirdly, it is clear that the measured effect is confounded with potential nonresponse bias. The respondents that could not have participated if the possibility to have the survey in their native language did not exist did show a more traditional attitude.
Despite the alternative explanations for the effect of interview language, the fact remains that it had a systematic effect. This means there is a real trade-off between cross-cultural comparability and reducing nonresponse among some ethnic groups.
As for the fourth research question – "How does interviewer gender and gender match affect the cross-cultural comparison?" – the results showed a significant effect for interviewer gender among Moroccans and gender match among Turkish when it came to attitudes towards *Gender roles.* Perhaps not surprisingly, female interviewers cause systematically less traditional attitudes towards *Gender roles* than male interviewers among

the Moroccans. Also, Turkish men have more traditional attitudes towards *Gender roles* when they are interviewed by a male interviewer compared to the Turkish men that were interviewed by a female interviewer. Turkish women are not systematically affected in their attitudes by the gender of the interviewer. In this case the hypothesis is partly confirmed. Interviewer gender and gender matching did effect the cross-cultural comparability, but the effect of interviewer gender was only discernible among Moroccan respondents and the effect of gender match was only present among Turkish male respondents.

With respect to the fifth research question – "How does the presence of others during the interview affect the cross-cultural comparison of attitudes on *Gender Roles* and *Family ties* between non-Western groups in the Netherlands?" – the results show that respondents of Surinamese and Antillean origin offered more traditional views in the presence of others. Among Surinamese respondents, this systematic effect was present on both factors, whereas for the Antilleans this only occurred for *Family ties.* Also in this instance the hypothesis is only partly confirmed. The presence of others during an interview resulted in more traditional views towards *Gender roles* and *Family ties*, but only among Surinamese and only with respect to *Family ties* among Antilleans.

The modeling of the incomparability of samples was done using a propensity score reweighting procedure of the sociodemographic variables that showed both a significant difference in the distribution between at least two ethnic groups and a significant effect on the indicators designed to measure the latent constructs.

The results for the sixth and final research question – "How much of the observed differences in attitudes on *Gender Roles* and *Family ties* between non-Western groups can be attributed to differences in sociodemographic composition between non-Western populations in the Netherlands?" – showed that the incomparability of samples explains some of the observed cross-cultural differences on both *Gender roles* and *Family ties*. In the case of *Gender roles*, this effect was large enough to render all observed differences between ethnic groups non-significant. This result confirms our sixth and final hypothesis that part of the observed differences between the ethnic groups can be explained by differences in sociodemographic composition.

It is important to be aware of the fact that survey data can be affected by a manifold of factors. These can be unwanted spin-offs of survey design choices or uncontrollable disturbance factors. In this case, it is clear that tailor-made response-enhancing measures and other, less controllable sources of method bias affect the cross-cultural comparison of non-Western minority ethnic groups, not only because they introduce a bias in estimates for an ethnic group, but, more importantly, because they impact the groups differently.

In the case of face-to-face surveys designed to compare ethnic groups or countries, these effects can lead to wrong conclusions about the relative positions of groups or countries. This can have serious consequences if the survey results contribute towards deciding whether or not a policy is effective in reducing an observed socioeconomic or sociocultural difference or if it informs the decision about the allocation of funds.

The comparability bias can be caused by differences in the size of the various sources of method bias that affects the groups or countries under investigation, by the differential impact of the same method bias between groups or by a combination thereof.

In the case of cross-cultural studies, it is important for the researchers to be aware of how the data were collected and how this can potentially bias survey estimates. This is especially important in the case of unexpected results based on data that used different data collection strategies among different ethnic groups.

With respect to data collected via face-to-face surveys it is recommended to take into account potential interviewer effects to avoid spurious effects, especially in the context of cross-cultural comparisons. In those cases when no information about the interviewer is available, one may consider using stricter criteria for significance testing, such as increasing the significance level to 0.01 instead of 0.05.

With respect to cross-cultural comparison, one also needs to consider how the research question is reflected by the results of the comparison. A substitution of observed differences between cultures with cultural differences is easily done, but that will mostly be confounded with differences in sociodemographic composition. For instance, observed differences in the *Gender roles* between the Turkish and Surinamese group can be interpreted as the average Turkish person being more traditional than the average Surinamese person. However, the average Turkish person has a different set of sociodemographic characteristics than the average Surinamese person. When Turkish and Surinamese persons with the same set of characteristics are compared the conclusion might be different.

The present study has several limitations that make the interpretation of the results not entirely straightforward. First of all, a MGCFA approach was used that included a cluster variable to adjust for interviewer-dependent correlation between the answers of respondents that were interviewed by the same interviewer. Given this modelling approach, it was not possible to compare the competing non-nested models using AIC or BIC fit indices. Therefore, the relative fit of the competing models was evaluated using fit measures that are not designed for comparing non-nested models and no conclusions could be drawn as to which of the models best describes the data. However, given the observed effects of the different sources of method bias on the cross-cultural comparability, we believe that we have adequately demonstrated the potential threat to making valid cross-cultural comparisons when these sources are not taken into account.

A second limitation concerns the quasi-experimental design used in this study. Data collected via this design does not allow for a complete disentanglement and entirely unbiased estimates of the different sources of identified method bias. Also, the data used in the present study did not allow for the complete disentanglement of the different ways (i.e., nonresponse, interview language and ethnicity) in which bilingual interviewers with a shared ethnic background can affect cross-cultural comparability.

A third limitation of the current study concerns the paradata. Several of the indicators measuring the existence of method bias are proxy estimates (i.e., interviewer assessments). A recommendation for further research could therefore be to include tape recordings of the interview in order to allow for more direct assessment of the effect of the interview language or of the extent to which others had an influence during (parts of) the interview.

As mentioned before, one can view the quasi-experimental design of this study as a drawback for this type of analysis. However, one should be aware of the fact that both the uncontrollable sources of method bias, such as the presence of others, as well as certain tailor-made response-enhancing measures are always confounded with sociodemographic characteristics of respondents in cross-cultural surveys. Therefore, one may wonder if one should put effort in designing a fully randomized experimental design to capture these effects. Instead it may be more interesting to attempt building a body of evidence based on data collected via more realistic quasi experimental designs such as the present one, in order to gain a better understanding of the effect these inherently confounded sources of method bias can have on the comparability of cross-cultural surveys and of the extent to which they can compromise cross-cultural comparisons. It might be preferable to collect more and/or more direct paradata and to further develop models that are better suited to correcting or testing for the existence of these effects based on data collected via quasi-experimental designs.

Appendices

Appendix 6.A

Parameter estimates and standard errors of the five multi group models

Parameter estimates (se)	M0	M1	M2	M3	M4
$\overline{Gr}_M$	0.106 (0.047)	0.085 (0.113)	0.049 (0.119)	0.202[a] (0.098)	0.096 (0.140)
Gr_S	0.556 (0.056)	0.595 (0.152)	0.530 (0.151)	0.831[a] (0.103)	0.513 (0.173)
Gr_A	0.511 (0.054)	0.516 (0.121)	0.479 (0.121)	0.743[a] (0.091)	0.479 (0.148)
$\overline{Ft}_M$	0.271 (0.053)	0.319 (0.084)	0.259 (0.093)	0.230[a] (0.066)	0.307 (0.084)
Ft_S	1.057 (0.066)	1.085 (0.094)	0.945 (0.097)	1.022[a] (0.065)	1.017 (0.103)
Ft_A	1.194 (0.069)	1.220 (0.087)	1.100 (0.093)	1.195[a] (0.055)	1.184 (0.101)
$Corr(Gr, FT)_T$	0.272 (0.029)	0.270 (0.039)	0.208 (0.038)	0.240 (0.041)	0.268 (0.027)
$Corr(Gr, FT)_M$	0.193 (0.029)	0.199 (0.041)	0.210 (0.046)	0.192 (0.047)	0.222 (0.048)
$Corr(Gr, FT)_S$	0.406 (0.047)	0.416 (0.103)	0.406 (0.102)	0.330 (0.080)	0.468 (0.155)
$Corr(Gr, FT)_A$	0.421 (0.045)	0.413 (0.073)	0.404 (0.075)	0.320 (0.057)	0.475 (0.082)
$\lambda^{Gr}_{Mangeld}$	1.000 (fixed)	1.000 (fixed)	1.000 (fixed)	1.000 (fixed)	1.000 (fixed)
$\lambda^{Gr}_{Inkjongs}$	0.951 (0.027)	0.949 (0.031)	0.949 (0.040)	0.956 (0.036)	0.949 (0.038)
$\lambda^{Gr}_{Vrwstopw}$	0.839 (0.025)	0.843 (0.036)	0.856 (0.042)	0.786 (0.042)	0.838 (0.037)
$\lambda^{Ft}_{Thuishuw}$	1.000 (fixed)	1.000 (fixed)	1.000 (fixed)	1.000 (fixed)	1.000 (fixed)
$\lambda^{Ft}_{Vertrfama}$	0.608 (0.033)	0.574 (0.040)	0.608 (0.045)	0.576 (0.060)	0.558 (0.039)
$\lambda^{Ft}_{Kibezoud}$	0.668 (0.034)	0.667 (0.047)	0.705 (0.059)	0.655 (0.073)	0.644 (0.050)
$\tau^{Mangeld}_1$	-1.419 (0.062)	-1.457 (0.146)	-1.550 (0.145)	-1.755 (0.332)	-1.608 (0.147)
$\tau^{Mangeld}_2$	-0.543 (0.042)	-0.528 (0.102)	-0.638 (0.100)	-0.774 (0.302)	-0.670 (0.105)
$\tau^{Mangeld}_3$	-0.079 (0.039)	-0.085 (0.097)	-0.160 (0.096)	-0.270 (0.292)	-0.176 (0.112)
$\tau^{Mangeld}_4$	1.003 (0.052)	1.017 (0.137)	0.966 (0.137)	0.868 (0.283)	0.983 (0.172)
$\tau^{Inkjongs}_1$	-1.371 (0.057)	-1.415 (0.131)	-1.471 (0.133)	-1.554 (0.322)	-1.501 (0.123)

Appendix 6.A (continued)

Parameter estimates (se)	M0	M1	M2	M3	M4
$\tau_2^{Inkjongs}$	-0.440 (0.039)	-0.434 (0.092)	-0.519 (0.093)	-0.516 (0.291)	-0.506 (0.099)
$\tau_3^{Inkjongs}$	-0.101 (0.038)	-0.120 (0.094)	-0.192 (0.092)	-0.148 (0.287)	-0.194 (0.107)
$\tau_4^{Inkjongs}$	0.981 (0.051)	1.006 (0.132)	0.965 (0.132)	1.031 (0.281)	0.965 (0.164)
$\tau_1^{Vrwstopw}$	-1.473 (0.059)	-1.457 (0.128)	-1.564 (0.127)	-1.451 (0.277)	-1.606 (0.123)
$\tau_2^{Vrwstopw}$	-0.573 (0.039)	-0.596 (0.081)	-0.714 (0.078)	-0.535 (0.251)	-0.715 (0.081)
$\tau_3^{Vrwstopw}$	-0.183 (0.035)	-0.208 (0.077)	-0.295 (0.076)	-0.131 (0.241)	-0.302 (0.084)
$\tau_4^{Vrwstopw}$	0.940 (0.047)	0.925 (0.126)	0.883 (0.130)	1.059 (0.242)	0.869 (0.145)
$\tau_1^{Thuishuw}$	-0.791 (0.051)	-0.722 (0.106)	-0.820 (0.125)	-0.692 (0.264)	-0.866 (0.099)
$\tau_2^{Thuishuw}$	0.315 (0.043)	0.335 (0.057)	0.201 (0.066)	0.416 (0.247)	0.235 (0.060)
$\tau_3^{Thuishuw}$	0.652 (0.047)	0.673 (0.058)	0.544 (0.066)	0.788 (0.249)	0.569 (0.066)
$\tau_4^{Thuishuw}$	1.825 (0.078)	1.827 (0.103)	1.697 (0.112)	1.995 (0.281)	1.838 (0.120)
$\tau_1^{Vertrfama}$	-0.606 (0.043)	-0.483 (0.076)	-0.545 (0.086)	-0.202 (0.187)	-0.646 (0.069)
$\tau_2^{Vertrfama}$	0.736 (0.040)	0.767 (0.046)	0.711 (0.054)	0.987 (0.208)	0.701 (0.051)
$\tau_3^{Vertrfama}$	1.432 (0.058)	1.408 (0.059)	1.370 (0.077)	1.625 (0.235)	1.411 (0.060)
$\tau_4^{Vertrfama}$	2.490 (0.098)	2.394 (0.107)	2.419 (0.142)	2.530 (0.301)	2.544 (0.099)
$\tau_1^{Kibezoud}$	-0.367 (0.039)	-0.297 (0.075)	-0.329 (0.080)	0.030 (0.206)	-0.375 (0.070)
$\tau_2^{Kibezoud}$	0.881 (0.044)	0.880 (0.066)	0.818 (0.072)	1.203 (0.233)	0.780 (0.067)
$\tau_3^{Kibezoud}$	1.286 (0.057)	1.266 (0.082)	1.200 (0.089)	1.600 (0.256)	1.165 (0.086)
$\tau_4^{Kibezoud}$	2.195 (0.090)	2.164 (0.139)	2.120 (0.152)	2.490 (0.325)	2.107 (0.149)
χ^2	552.900	302.735	285.621	475.207	273.996
Df	92	92	92	348	92

Note. GR= Gender Roles and FT= Family Ties; T= Turkish; M= Moroccans; S=Surinamese; A= Antilleans; $GR_{Turkish}$ and $FT_{Turkish}$ are both set to zero. λ^{factor} = factorloading of the indicator; τ_x = threshold value of the indicator. [a] = adjusted for the (different) impact of the presence of others, own language, interviewer gender and gender match interaction between ethnic groups.

Appendix 6.B
Multiple causes results for Model 3 for Gender Roles (GR) and Family Ties (FT), separately for each ethnic group

Parameter estimates (se)	Turkish (N=812)		Moroccans (N=805)	
	GR	FT	GR	FT
Intercept	0.000 (0.000)	0.000 (0.000)	0.566 (0.371)	0.583 (0.432)
Big City Dweller	-0.340 (0.183)	-0.069 (0.098)	-0.198 (0.128)	-0.154 (0.141)
Employed	0.229 (0.075)*	0.018 (0.073)	0.179 (0.066)*	0.019 (0.177)
Has Child(ren)	-0.180 (0.171)	-0.388 (0.114)*	0.110 (0.105)	0.019 (0.177)
Has a partner	0.050 (0.093)	-0.096 (0.089)	-0.120 (0.092)	-0.260 (0.140)
Educational level	0.101 (0.043)*	0.180 (0.046)*	0.082 (0.036)*	0.104 (0.047)*
Male	-0.232 (0.090)*	0.176 (0.101)	-0.579 (0.081)*	-0.217 (0.113)
First generation immigrant	0.032 (0.154)	0.013 (0.121)	0.093 (0.125)	-0.192 (0.122)
Age group (ref group is 15-24)				
25 – 34 year	0.046 (0.149)	0.443 (0.167)*	0.126 (0.104)	0.288 (0.162)
35 – 44 year	0.162 (0.174)	0.558 (0.178)*	-0.013 (0.140)	0.353 (0.288)
45 – 54 year	0.016 (0.189)	0.554 (0.191)*	0.004 (0.145)	0.495 (0.211)*
55 – 64 year	0.069 (0.139)	0.510 (0.179)*	-0.045 (0.182)	0.513 (0.286)
65 + year	-0.075 (0.221)	0.344 (0.232)	-0.115 (0.183)	0.331 (0.262)
Others were present	-0.249 (0.160)	-0.109 (0.170)	-0.012 (0.159)	-0.298 (0.184)
Interviewed in native language	-0.142 (0.100)	-0.364 (0.131)*	-0.105 (0.114)	-0.356 (0.140)*
Gender match interaction	-0.294 (0.117)*	-0.048 (0.162)	0.133 (0.184)	0.015 (0.208)
Gender interviewer	-0.022 (0.157)	-0.043 (0.156)	-0.339 (0.159)*	-0.070 (0.195)

Note. * = p <0.05

Appendix 6.C: Observed differences on sociodemographic variables between ethnic groups after weighting for population distribution (Table C1) and after propensity score weighting (Table C2).

Surinamese (N=779)		Antilleans (N=852)	
GR	FT	GR	FT
0.404 (0.485)	1.272 (0.388)*	0.611 (0.362)	1.395 (0.348)*
0.045 (0.141)	-0.113 (0.100)	-0.219 (0.092)*	-0.273 (0.120)*
0.182 (0.109)	0.101 (0.084)	0.044 (0.068)	0.059 (0.079)
0.080 (0.093)	-0.071 (0.097)	0.008 (0.111)	-0.225 (0.096)*
0.088 (0.067)	0.085 (0.073)	0.064 (0.071)	0.077 (0.073)
0.171 (0.065)*	0.088 (0.048)	0.187 (0.047)*	0.272 (0.052)*
-0.671 (0.145)*	-0.057 (0.074)	-0.604 (0.099)*	-0.080 (0.099)
-0.223 (0.093)*	-0.426 (0.096)*	-0.286 (0.094)*	-0.264 (0.113)*
0.004 (0.139)	0.168 (0.130)	0.004 (0.105)	-0.090 (0.119)
-0.153 (0.131)	0.221 (0.151)	0.017 (0.116)	-0.007 (0.150)
-0.115 (0.149)	0.106 (0.137)	0.075 (0.140)	0.130 (0.146)
-0.066 (0.154)	0.133 (0.152)	0.001 (0.127)	-0.219 (0.159)
-0.147 (0.183)	-0.061 (0.161)	-0.135 (0.182)	-0.060 (0.229)
-0.689 (0.202)*	-0.405 (0.162)*	-0.062 (0.119)	-0.259 (0.100)*
-0.414 (0.856)	-0.013 (0.438)	-0.169 (0.132)	-0.241 (0.113)*
0.168 (0.182)	0.123 (0.117)	0.006 (0.126)	-0.079 (0.145)
0.031 (0.197)	-0.050 (0.120)	-0.054 (0.102)	-0.094 (0.128)

Table C1
Observed differences on sociodemographic variables between ethnic groups after weighting for population distribution.

				Significant differences between ethnic groups (bonferonni adjusted)		
Variable	Ethnic group	estimate	se	Turkish	Moroccans	Surinamese
Men (proportion)	Turkish	0.517	0.019			
	Moroccans	0.506	0.018			
	Surinamese	0.464	0.018			
	Antilleans	0.494	0.018			
Age Group (mean)	Turkish	2.750	0.052			
	Moroccans	2.739	0.053			
	Surinamese	3.079	0.054	*	*	
	Antilleans	2.710	0.052			*
First generation immigrant (proportion)	Turkish	0.693	0.018			
	Moroccans	0.664	0.017			
	Surinamese	0.646	0.017			
	Antilleans	0.721	0.016			*
Educational level (mean)	Turkish	2.074	0.039			
	Moroccans	2.005	0.038			
	Surinamese	2.607	0.037	*	*	
	Antilleans	2.533	0.035	*	*	
Big City Dweller (proportion)	Turkish	0.228	0.016			
	Moroccans	0.299	0.016	*		
	Surinamese	0.360	0.018	*		
	Antilleans	0.254	0.016			*
Employed (proportion)	Turkish	0.489	0.019			
	Moroccans	0.488	0.018			
	Surinamese	0.674	0.017	*	*	
	Antilleans	0.601	0.018	*	*	*
Has child(ren) (proportion)	Turkish	0.632	0.019			
	Moroccans	0.591	0.018			
	Surinamese	0.615	0.018			
	Antilleans	0.548	0.018	*		
Has partner (proportion)	Turkish	0.579	0.019			
	Moroccans	0.573	0.018			
	Surinamese	0.506	0.018	*	*	
	Antilleans	0.458	0.017	*	*	

Note. * p<0.05/no. of pairwise comparisons. Variables included in the population weights: gender, household size, municipality size, immigration generation, age groups (12)

7 Summary and conclusion

The demand for information about ethnic minorities, originating mostly from national, but also from supranational government and institutions, doesn't seem to be diminishing. For example, policy objectives continue to require the collection of information on themes like the socioeconomic position and the degree of sociocultural integration of ethnic minorities to evaluate them. A part of this information can only be collected by means of surveys, despite the availability of public records and administrative registers. In order to be able to make full use of the surveys' potential to monitor and evaluate policy, it is essential that the data should render an accurate picture of the situation of ethnic minorities. In other words: there should be no doubt about the quality of the data and its fitness for purpose for the research question in point.

However, obtaining accurate survey data about ethnic minorities is a challenge, due to problems regarding the representation of ethnic minorities in surveys and also to measurement issues. Both reasons can cause reduced accuracy of sample estimates. Furthermore, accuracy is only one dimension that needs attention in the process of evaluating data quality. Other important dimensions are the relevance of the data, the timeliness with which it becomes available, the accessibility and clarity of the data and its coherence and comparability.

The difficulties associated with collecting data among ethnic minorities and with possible measures taken in order to obtain a better representation and measurement of and among ethnic minorities have, in turn, consequences on the other quality dimensions. For instance, comparison or benchmarking is one of the important reasons for collecting data. Both the way in which data on ethnic minorities has been collected and the level to which it reflects the population under study can affect the comparability of this data with data collected about other groups, with data collected in other waves, or with data collected among the general population. In case of a biased comparison, it is easy to draw wrong conclusions about the success of a policy measure.

The present study set out to investigate the quality of survey data collected among non-Western minorities in the Netherlands and how this might relate to the survey design. In order to answer the research question regarding the quality of survey data and its relation to the survey design, we mainly looked at the two quality dimensions that seem most pertinent for data about ethnic minorities: *accuracy* and *comparability*. We focused on two aspects of *accuracy*: 1) representation, that is, how well the population is reflected by the respondents to the survey and 2) measurement, that is, to what degree the manner of administering the survey affects the measurement of the substantive topics among respondents to the survey. With respect to *comparability,* we focused on how comparable the survey data collected between different minority groups are.

This last chapter intends to summarize the previous chapters and to review the most important conclusions of the present study with regard to the research question. In this chapter, we also address the relevance of this thesis for researchers within and outside of the Netherlands. In the end, we shall look at the results of this study from a different perspective and focus on the relationship among research objective, data quality, and fitness-for-purpose.

In chapter 1 we introduced the study and outline the thesis. Chapter 2 looked at the difficulties concerning the definition of ethnicity and ethnic minorities and their consequences. It also provided an overview of the literature concerning the problems that can arise when conducting surveys among ethnic minorities. These problems were then correlated to specific *Total Survey Error* sources within the representation and the measurement dimensions. Furthermore, chapter two provided an overview of the literature on the measures that can be taken in order to ensure a better representation of ethnic minorities in surveys and it discussed ways of assessing the success of such measures. Attention was also paid to the trade-off: to what degree might the measures taken to ensure a better representation of the population affect the measurement of substantive variables among the respondents. The last section of chapter two approached the issues of comparability and timeliness of data collected among ethnic minorities. The focus was placed on the ways in which survey design choices may negatively influence quality in terms of comparability and timeliness of the data collected among ethnic minorities. Finally, cost-related considerations were presented in connection to different survey designs and it was discussed how they should be included in the trade-off between quality and cost of data collection among ethnic minorities.

Chapters 3 to 6 were (quasi)-experimental studies. As far as the accuracy of the data is concerned, we analyzed the degree to which the survey design affects how well the population is reflected by the respondents to the survey *and* to what degree the manner of administering the survey affects the measurement of the substantive topics among respondents. As far as comparability is concerned, we studied the degree in which different survey design choices and factors beyond the control of the survey designer can influence the comparability of the survey data collected between different minority groups. The studies also assessed how financial and time restrictions affect data collection and, subsequently, the quality of survey data collected among ethnic minorities.
The data used in the (quasi)-experimental chapters of this study comes from the SIM monitor. The SIM monitor is conducted on behalf of the Netherlands Institute for Social Research/SCP. The survey sponsor is the directorate of Knowledge and Prospective study in the Ministry of Infrastructure and the Environment that is responsible for Housing, Neighbourhoods and Integration. The Survey on the Integration of Minorities (SIM) sets out to measure the socioeconomic position of the four largest non-Western minority populations living in the Netherlands and an autochthone Dutch control group, as well as their sociocultural integration. These four groups are Dutch of Turkish descent, Dutch of Moroccan descent, Dutch of Surinamese descent and Dutch of Antillean or Aruban descent.
The SIM is a nationwide, cross-sectional survey, which started in 2006 and was repeated in 2011. In 2006, Statistics Netherlands drew a random probability sample from each of the four largest non-Western minority populations living in the Netherlands and an autochthonous Dutch control group. Subsequently, the data was collected by means

of face-to-face computer-assisted personal interviewing (CAPI). Several tailor-made response-enhancing measures were used, such as the use of translated questionnaires, bilingual interviewers and interviewers with a shared ethnic background. The number and intensity of the tailor-made response-enhancing measures varied between the four ethnic groups.

The wave of 2011 also comprises a large-scale survey design experiment. Statistics Netherlands drew two random probability samples from each of the four largest non-Western minority populations living in the Netherlands and an autochthon Dutch control group. Subsequently, one sample was assigned to a face-to-face CAPI design, while the other sample was assigned to a sequential mixed-mode design (MM) using computer-assisted web interviewing (WEB), computer-assisted telephone interviewing (CATI) and face-to-face CAPI. Both 2011 survey designs involved the use of tailor-made response-enhancing measures, such as the use of translated questionnaires, bilingual interviewers and interviewers with a shared ethnic background. The number and intensity of the tailor-made response-enhancing measures varied between the four ethnic groups, but not within each ethnic group on both arms of the experiment. The thesis focused on survey design and survey data quality among non-Western minorities in the Netherlands. This is why the samples containing autochthonous Dutch are excluded from this thesis.

In total, data from 12 different sub-surveys were used in this study: data from three different surveys -SIM 2006 CAPI; SIM 2011 CAPI; SIM 2011 MM- among each of the four largest non-Western minority populations living in the Netherlands.

The main focus of chapters 3 and 4 was on the representation of ethnic minorities in surveys. We tried to determine whether there was a definite relation between different survey designs and survey design choices and the representation of the four non-Western ethnic minority populations in the respondent sample (i.e., those who completed the survey). Chapter 3 investigated how different survey design choices – differences in the duration and timing of the fieldwork, the use of bilingual interviewers with the same ethnic background and the use of a reissue – affect the composition of the respondent sample and how this might relate to the occurrence of nonresponse bias on survey estimates in surveys conducted among non-Western minorities in the Netherlands. Data from a total of eight sub-surveys were used in this study: four from both SIM2006 CAPI and SIM2011 CAPI.

In Chapter 4 investigated the effect that the use of different data collection methods in surveys may have on the representation of the four minority populations in the respondent samples. We studied the way in which the use of a sequential mixed-mode design in surveys among non-Western minorities in the Netherlands affects the quality of the respondent sample compared to a single-mode face-to-face design, and how these two designs can potentially impact nonresponse bias. A second point of interest was whether these designs systematically enhance response rates differently among various sociodemographic subgroups among non-Western minorities. Finally, costs and cost-related issues particular to this sequential mixed-mode design that are relevant in the quality versus costs trade-off decision were taken into consideration. Data from a total of eight sub-surveys were used in this study: four from both SIM2011 CAPI and SIM2011 MM.

For the analyses in Chapter 3 and 4, we used different quality indicators concerning representation in addition to the response rate: the representativity-indicator or R-indicator, partial r-indicators, the standardized maximal absolute bias, fieldwork disposition codes and fraction of missing information (FMI). The conclusions we can draw about the representation of the target populations on the basis of this combination of indicators are different from the conclusions one would draw if response rate were the only quality indicator used. For instance, it turns out that design choices that do not adapt to the social reality of the non-Western ethnic minorities in the Netherlands do not only lead to additional nonresponse, but also to more selective nonresponse. Furthermore, the use of a sequential mixed-mode survey design (Web-CATI-CAPI) did generate a higher response, but it also led to less representative respondent samples and showed more potential for nonresponse bias in survey estimates than a single-mode CAPI survey design. As a result, the studies in these chapters offer a different perspective on the relation between survey design choices and the representation of non-Western ethnic minorities in the Netherlands.

When it comes to evaluating the effect of separate response-enhancing measures in surveys, it is important to note that, in many circumstances, analysis methods such as logistic regression may give biased results because of the non-random allocation of sample units to ‘treatments’. For example, persuasion letters are only sent to reluctant respondents, and therefore seem to have a negative effect on response rates as reluctant respondents more often turn into final refusers and no persuasion letters are sent to respondents who cooperate instantaneously. Differences in response rate are therefore not really informative, as opposed to differences in the final sociodemographic composition of the respondent group or differences in the potential for nonresponse bias.

Another important result was that the potential cost savings obtained by introducing cheaper modes came at the expense of the loss of quality in terms of representativity and potential nonresponse bias. Furthermore, these theoretic savings were calculated based on a biased view about the actual costs that ignored other relevant factors. For example, compared to a face-to-face survey design, the use of a sequential mixed-mode design limits the duration of the interview. In this study the WEB and CATI questionnaire were about two-thirds of the length of the CAPI questionnaire. This means a substantial loss of information, which in fact translates to an increase of the cost per survey question in the sequential mixed-mode survey. Additionally, each mode needs a certain amount of time to be used to its full potential before switching to the next mode. This can increase the length of the fieldwork period, which means later data delivery. Moreover, it should not be forgotten that a sequential mixed-mode design will require additional time for adapting questionnaires to different modes. Extra time and resources also have to be dedicated to checking and correcting for potential mode effects that can distort the results.

Especially for relatively small sample sizes and known survey difficulties in connection with specific ethnic minority target populations, these additional costs and extra time may compensate the expected savings. In this particular study it was concluded that the actual cost savings did not outweigh the reduction in quality and content which has a direct impact on the relevance of the data.

At the same time, Chapters 3 and 4 provided insight into the ways in which approach strategies and design choices can be adapted in surveys among non-Western minorities in the Netherlands in order to obtain a better balanced sociodemographic composition of the respondent group. In our case, given the survey design, it is advisable to avoid using CATI in a mixed-mode survey because it leads to a small and selective group of respondents. However, introducing a reissue of nonrespondents to another interviewer at an earlier stage, instead of extended calls in the first phase might be a successful measure, especially in the case of these particular populations.

Chapter 5 focused on how the manner of administering the survey affects the measurement of the substantive topics among ethnic minority respondents. This chapter described an experimental study that investigated the impact of different modes in conjunction with tailor-made response-enhancing measures on the measurement of ten substantive variables in surveys among four non-Western minorities in the Netherlands. Specifically, we studied the extent to which the use of different modes together with tailor-made response-enhancing measures elicited measurement differences among respondents. For the analyses we used a recently developed method for disentangling the impact that a data collection method has on the potential respondents' choice to participate (selection-effect) and the impact that a data collection method has on the answers provided by a respondent (mode-effect). To assist in the interpretation of the results we used a 'pattern' approach. This meant using different patterns observed across the separate ethnic group results to facilitate the identification of measurement and selection effects. This approach allowed us to establish whether observed measurement and selection effects are the result of mode, of tailor-made response-enhancing measures, of violations of the assumptions underlying this recently developed method, or of a combination thereof. Data from a total of eight sub-surveys were used in this study: four from SIM2011 CAPI and four from SIM2011 MM.

The results showed rather consistent measurement effects for seven out of ten variables. Measurement effects occur more often on *sociocultural* questions, but also, occasionally, on more *sociostructural* or *background* questions. Furthermore, these effects are found despite the fact that extensive efforts were undertaken to minimize mode effects and translation effects. Data collected in surveys that do not undertake these extensive measures is likely to suffer far more from unwanted measurement effects.

With respect to how the manner of administering the survey affects the measurement of the substantive topics among ethnic minority respondents, it was clearly demonstrated that the use of multiple modes *in combination with* tailor-made response-enhancing measures did increase measurement variability compared to the single-mode with tailor-made response-enhancing measures. Furthermore, web interviews seem to elicit less socially desirable responses compared to when interviewers are used, mainly in the case of sociocultural oriented questions. However, some tailor-made response-enhancing measures, such as the use of *bilingual* interviewers with a shared ethnic background remain necessary among populations with language barriers. Among populations without significant language barriers, the benefits of using interviewers with a shared ethnic background are more difficult to assess. Still, accurate measurement is only *one* quality issue. It is important to consider that the comparability of the different ethnic groups might actually be reduced if the interviewers with shared ethnic background are only used in certain ethnic (sub)groups.

Chapter 6 addressed the question of comparability of survey data collected among different ethnic minority groups. A study was conducted to assess the impact of several sources of method bias on the cross-cultural comparison of survey outcomes among four non-Western minority ethnic groups living in the Netherlands. In particular, we investigated how interviewer effects, the use of an interviewer with a shared ethnic background, interview language, interviewer gender, gender matching, the presence of others during the interview and differences in the sociodemographic sample composition affected the cross-cultural comparison of attitudes towards gender roles and family ties between these groups. Data from a total of four sub-surveys from the SIM2011 CAPI were used in this study. For the analysis we introduced methods to estimate the potential impact of method bias on cross-cultural comparisons.
The results show that measurements of gender roles and of family ties constructs were full scalar invariant across the different ethnic groups, but that observed differences in attitudes between ethnic groups, especially towards gender roles, were affected by method bias. Interviewer effects as well as interview language, interviewer characteristics (i.e., ethnicity and gender), gender matching, the presence of others during the interview and differences in the sociodemographic sample composition all affected the measurement of at least one of the constructs in at least one of the ethnic groups. This led to biased comparisons which were the result of differences in the size of the various sources of method bias that affected the groups, of the differential impact of the same method bias between groups or of a combination thereof.

On the relevance of the thesis

As mentioned earlier, we wish to dedicate some attention to the relevance of the results of the present research. Within the Netherlands, this study is relevant for survey sponsors and researchers studying non-Western minorities, but also to end-users of minority data. The four non-Western minority groups used in this study account for about two-thirds of the total population of non-Western minorities living in the Netherlands. As for the monitor, different survey design choices have been made for the SIM surveys: between waves *and* between groups. This enabled us to assess both the effect of different survey designs on representation and measurement and the effect on the comparability of the data.
The development of a methodology for highlighting the systematic relation between survey design (choices) and the measurement and representation of these minority groups offers insight into how accurate a particular survey is in rendering a picture of these groups. This, in turn, provides an instrument for researchers and other users of minority data to determine whether the data collected among these groups – in which their social reality has been taken into account to a higher or a lesser degree – is well fit for answering a research or a policy question. This could be used, for instance, when the government or the political institutions are interested in the effect of minority policies on the socioeconomic position or sociocultural integration of non-Western minorities

in the Netherlands. The study also offers suggestions of ways in which the measurement and representation of these groups could be improved and points out aspects worth paying attention to when comparing groups.

We can point out several reasons why the results of this study are also relevant for a wider public, for instance for ethnic minority survey research in other countries or for users of survey data collected among ethnic minorities from different countries. Firstly, these results might be interesting and relevant for a wider (international) public *precisely* because of the challenges posed by the chosen ethnic minority groups, the different survey design choices and the tailor-made measures. Together, these four groups illustrate a wide range of the difficulties and challenges experienced at international level. In this particular pool we were confronted with, among others, linguistic problems, larger cultural differences in relation to the main society, a significant degree of stigmatization, a substantial incidence of functional illiteracy, the lack of a written language and a higher degree of sociodemographic correlates that contribute to nonresponse or insufficient coverage in ethnic minority groups. The different survey design choices and the tailor-made measures are, in turn, rather universal in their applicability, while, consequently, the results showing a systematic effect of these choices and measures on the measurement and representation can also be interpreted that way.

A second reason is to be found in the position of the Netherlands in survey research among ethnic minorities. The Netherlands has a long tradition of migration and of survey research being conducted among migrants and ethnic minorities for policy purposes in which ethnic minorities are classified according to the post-migration multicultural classification strand. That is, ethnicity is substituted with migrant origin, so as to include not only recent immigrants, but also their descendants. Some countries, especially in North-Western Europe, research ethnic minorities on account of similar policy objectives and use a similar classification, therefore the results could be relevant to them. Other countries, particularly those with a shorter experience of immigration, might be taking this step in the near future. Furthermore, the samples used in this study were drawn using a fairly complete sampling frame with additional information about the sampled persons which may not be available in other countries. During the data collection stage additional paradata, that is process data and interviewer observations, was also recorded which may not be allowed or possible in other countries. The availability of all this paradata on respondents and nonrespondents allowed for a more detailed analysis of the effect of survey design (choices) and other factors on the measurement and representation. What's more, the study of the impact of tailor-made response-enhancing measures, such as translated questionnaires and bilingual interviewers with a shared ethnic background, on the representation and measurement can be useful for researchers studying indigenous ethnic minorities as well.

A third reason for relevance is the shift in focus as far as the quality of survey data in these hard to research populations is concerned. In this sense, the fact that we looked at more than just accuracy when assessing the data quality is relevant because of the increasing diversification of ethnic minority groups in the countries. This also bears relevance on the comparison of different ethnic minority groups between countries. For example, one can identify quite substantial differences among the chosen groups in

terms of sociodemographic composition, cultural distance to the Dutch society, reason for immigration etc., which resonate with the heterogeneity of ethnic minority groups observed in other countries.
A fourth reason for relevance is the application of different recently developed quality indicators. This adds insight into possible analysis strategies that can be used to assess and possibly improve survey research among ethnic minorities.

The results of the different chapters make it clear once again that the research objective and the survey design are important for assessing the fitness for purpose of the data, especially when this data is re-used in order to inform on other research questions. This can generate a shift in the importance allotted to the different quality dimensions. Moreover, the explicit review of cost-related factors should make people more aware of the concessions made when choosing for alternatives that appear cheaper at first sight. We also hope to have clarified that the fitness for purpose of the data collected in surveys among ethnic minorities or among the general population comprising ethnic minorities in which the design and execution of the survey did not take into account the social reality of ethnic minorities should be assessed critically – especially in situations where one needs to report specifically on different ethnic minority (sub)groups. Needless to say the same critical approach is needed in assessing surveys in which the social reality of ethnic minorities has been taken into account. In these cases one may encounter not only a situation of trade-off between measurement and representation as a result of tailor-made response-enhancing measures, but also a similar incompatibility between the accuracy and the comparability of the data.
It is our hope that the present study accomplished more than increasing awareness about the impact of survey design choices and the necessity of collecting data about the data collection process itself in the case of ethnic minority research. We hope that the choices made in designing surveys and collecting data among ethnic minorities will be documented in more detail from now on. Can we actually afford not to?

Samenvatting

De vraag naar informatie over etnische minderheden is onverminderd groot.
Een belangrijke reden is dat nationale en regionale overheden informatie nodig hebben over de socio-economische positie en de mate van integratie van etnische minderheden om beter beleid te kunnen voeren. Registraties kunnen een deel van deze informatie leveren, maar over veel onderwerpen, zoals religiositeit, identiteit en integratie zijn geen registraties beschikbaar. Daarom blijven enquêtes nodig. Het is van groot belang dat de resultaten van enquêtes van een goede kwaliteit zijn, dat ze geschikt zijn om de onderzoeksvragen te beantwoorden, kortom: dat men op de resultaten van deze enquêtes kan vertrouwen.
Het verkrijgen van accurate enquêtegegevens over etnische minderheden is niet simpel. Etnische minderheden zijn meestal ondervertegenwoordigd in enquêtes. Bovendien is het maar de vraag of de respondenten die wel meedoen aan de enquête representatief zijn voor de groep waarin we geïnteresseerd zijn. Doen bijvoorbeeld alleen mensen met een hoge opleiding mee, of mensen met een goede baan, of mensen die in Nederland geboren zijn? Om ervoor te zorgen dat iedereen mee kan doen is maatwerk nodig, zoals het vertalen van vragenlijsten en de inzet van interviewers met dezelfde etnische herkomst als de beoogde respondent. Dit maatwerk kan leiden tot een hogere respons, maar meet je met een vertaalde vragenlijst hetzelfde als met de oorspronkelijke? Als je op de resultaten van enquêtegegevens wil kunnen vertrouwen, moeten de gegevens van goede kwaliteit zijn, moet je verschillende groepen goed kunnen vergelijken – ondanks taalverschillen – en moeten de gegevens ook liefst snel beschikbaar zijn. Hoe snel je over de resultaten wil beschikken, en hoeveel budget je hebt, stelt natuurlijk ook weer een bovengrens aan de kwaliteit: een hoge respons vereist bijvoorbeeld een langere veldwerkperiode.
De relatie tussen de opzet van een enquête en de kwaliteit van de enquêtegegevens over niet-Westerse minderheden in Nederland staat centraal in deze dissertatie. Hierbij is voornamelijk gekeken naar twee aspecten van datakwaliteit die hier het meest relevant lijken: *accuraatheid* en *vergelijkbaarheid*. Voor de *accuraatheid* is vooral gekeken naar 1) representatie, dat wil zeggen, hoe goed is de onderzoekspopulatie vertegenwoordigd door de respondenten van een enquête en 2) meting, dat wil zeggen, hoe is de manier waarop de gegevens zijn verzameld bij de respondenten van invloed op de antwoorden. Met vergelijkbaarheid staat hier de vergelijkbaarheid van gegevens van verschillende minderheidsgroepen centraal. Ook is er in deze dissertatie aandacht besteed aan de kosten van enquêtes onder niet-Westerse minderheden.
Hoofdstuk 1 beschrijft het onderwerp en de opzet van deze dissertatie. Hoofdstuk 2 verschaft het theoretisch kader. Hier is o.a. aandacht besteed aan de vraag hoe we termen als etniciteit en etnische minderheid kunnen definiëren en operationaliseren. Verder wordt een overzicht gepresenteerd van de internationale literatuur over het enquêteren van etnische minderheden.

Hoofdstukken 3 t/m 6 doen verslag van (quasi)-experimentele studies. In hoofdstukken 3 en 4 staat de vraag naar representatie centraal, ofwel: vormen de respondenten een goede vertegenwoordiging van de groep die we willen onderzoeken? Er is onderzocht of verschillen in enquêteopzet systematisch van invloed zijn op wie er mee doet en wie niet. Bij de opzet van een enquête is het onder andere van belang wanneer een onderzoek het veld in gaat, en hoeveel tijd er is voor de dataverzameling. Ook de manier waarop data worden verzameld kan invloed hebben: door interviewers aan huis, telefonisch of via een web-enquête. In hoofdstuk 3 zijn twee enquêtes met elkaar vergeleken die verschillen in de mate waarin ze rekening houden met de leefsituatie van niet-Westerse minderheden. Aansluiten bij de leefsituatie behelst bijvoorbeeld het aanbieden van vertaalde vragenlijsten, de inzet van tweetalige interviewers met dezelfde etnische afkomst, of kortere veldwerkperiodes omdat onder niet-Westerse minderheden het vaker voorkomt dat men moeite heeft met de Nederlandse taal, dat met name oudere, niet-Westerse minderheden vaker functioneel analfabeet zijn en dat men vaker verhuist. In hoofdstuk 4 zijn twee enquêtes met elkaar vergeleken die verschillen in de manier waarop de data wordt verzameld. De ene enquête maakt uitsluitend gebruik van interviewers aan de deur in combinatie met maatwerk zoals de inzet van vertaalde vragenlijsten en de inzet van tweetalige interviewers met dezelfde etnische afkomst (interviewer-aanpak). Bij de andere enquête kon men de vragenlijst eerst op het internet beantwoorden. Als men niet meedeed werd men daarna opgebeld voor een telefonisch interview en als dat geen succes opleverde kwam de interviewer aan de deur. Deze laatste enquête maakte gebruik van hetzelfde maatwerk en de veldwerkperiode was even lang (de combi-aanpak).

Op basis van een breed scala van analysemethoden kunnen de volgende conclusies getrokken worden. Enquêtes die meer aansluiten bij de leefsituatie van niet-Westerse minderheden leveren een representatiever beeld van de doelgroep op dan een meer standaardaanpak. Enquêtes die meer aansluiten bij de leefsituatie zijn dus niet zozeer gericht op het verkrijgen van een hogere respons, maar zorgen er wel voor dat de respons beter verdeeld is over verschillende subgroepen waarin we geïnteresseerd zijn: iedereen kan meedoen. Met de inzet van tweetalige interviewers met dezelfde etnische achtergrond verbetert de representatie van de niet-Westerse minderheidsbevolking in Nederland: ook de eerste generatie migranten is dan beter vertegenwoordigd. Het blijkt verder dat de combi-aanpak in vergelijking met de interviewer-aanpak wel leidt tot een hogere respons, maar niet tot een meer evenwichtige vertegenwoordiging van de doelgroep, zelfs als vertaalde vragenlijsten worden gebruikt en in de laatste fase aan de deur interviewers uit dezelfde groep worden ingezet

In hoofdstuk 5 is onderzocht of de manier waarop de enquêtevraag wordt afgenomen leidt tot antwoordeneffecten bij respondenten. Er is bij de combi-aanpak met vertaalde vragenlijsten en tweetalige interviewers uit de eigen groep onderzocht of de manier waarop de enquête wordt afgenomen invloed heeft op de antwoorden. Er blijken systematische verschillen te zijn afhankelijk van de manier van afname, waarbij web-enquêtes leiden tot de eerlijkste antwoorden onder respondenten. Hier is dus sprake van een dilemma: de inzet van tweetalige interviewers aan de deur met dezelfde etnische achtergrond is noodzakelijk om een representatief beeld te krijgen van etnische minderheidspopulaties waar sprake is van een hogere mate van taalproblemen en/of functionele ongeletterdheid, terwijl deze aanpak meer sociaal-wenselijke antwoorden uitlokt.

In hoofdstuk 6 staat de vraag naar de vergelijkbaarheid van enquêtegegevens centraal. In dit hoofdstuk is eerst onderzocht in welke mate ongewenste, methodologische meeteffecten, zoals de etnische achtergrond of het geslacht van de interviewer, of de taal waarin het interview is afgenomen, van invloed zijn op de antwoorden van de respondenten. De vraag was hierbij of bij iedere etnische groep deze ongewenste meeteffecten niet of in gelijke mate van invloed waren, of dat deze invloed tussen de groepen verschilden. In het eerste geval is de vergelijkbaarheid van gegevens tussen groepen niet in het geding, maar in het tweede geval is de vergelijking niet zuiver. Uit de analyses bleek dat er sprake was van een onzuivere vergelijking tussen groepen. De antwoorden van sommige etnische groepen werden meer beïnvloed door ongewenste methodologische meeteffecten dan die van andere groepen. Zo bleken Turkse mannen traditioneler te gaan antwoorden wanneer ze werden bevraagd door een mannelijke interviewer van Turkse afkomst, terwijl dit effect niet optrad bij Turkse vrouwen of bij een van de andere etnische groepen. Het gevolg hiervan is dat in een onderlinge vergelijking de Turkse mannen gemiddeld genomen traditioneler worden gezien terwijl dit deels een methodologisch artefact is.

In het 7^{e} en laatste hoofdstuk zijn alle belangrijke uitkomsten met betrekking tot de hoofdvraag van deze dissertatie nog eens op een rij gezet. Voorts is er ingegaan op de vragen voor wie en waarom de resultaten van deze dissertatie relevant zijn.

Summary

The demand for information about ethnic minorities remains at a constantly high level. An important reason is that national and regional administrations need information about the socioeconomic position and the degree of integration of ethnic minorities in order to implement better policies. Official registers can provide a part of this information, but no data is available in the registers on topics like religious beliefs, identity and integration. As a consequence, surveys remain necessary. It is very important that the results of these surveys should be of a high quality and that they should be fit for answering the research questions, in other words that the results of such surveys can be trusted. Obtaining accurate survey data about ethnic minorities is not easy. Ethnic minorities are usually underrepresented in surveys. Furthermore, it is not certain that people that do take part in surveys are representative of the group one is interested in. Do, for instance, only people with high education take part, or people with a good job, or people who were born in the Netherlands? To ensure that everybody can participate, one needs a tailor-made approach, such as translating the questionnaires and using interviewers with the same ethnic background as the intended respondents. This tailor-made approach can lead to higher response, but does a translated questionnaire still measure the same things as the original? In order to be able to trust the results, survey data needs to be of good quality, different groups need to be readily comparable – in spite of linguistic differences – and the data should also, preferably, be available quickly. How quickly one wishes to obtain the results and the overall budget also determine the upper limit of the data quality: to get a high response, for instance, one needs a longer fieldwork period.

The relation between the survey design and the quality of the survey data related to non-Western minorities in the Netherlands is the main focus of this dissertation.

With respect to data quality, the focus was on two aspects of data quality that seemed the most relevant in this context: *accuracy* and *comparability*. For accuracy, we studied mainly 1) representation, or how well the population under study is represented by the respondents of a survey, and 2) measurement, meaning how the manner in which data has been collected among respondents may affect the answers they provide. With respect to comparability, the focus has been on the comparability of data collected from different minority groups. Attention has also been paid in the dissertation to the costs of surveys among non-Western minorities.

Chapter 1 describes the subject and the structure of the thesis. Chapter 2 offers the theoretical framework. Among other things, we looked here at the question of how terms like ethnicity and ethnic minority can be defined and operationalized. A review is also provided of the international literature on surveying ethnic minorities.

Chapters 3 to 6 report on (quasi)experimental studies. Chapters 3 and 4 focus on the issue of representation, testing how well the respondents represent the group that we wish to investigate. We checked whether differences in survey design have a systematic influence on who takes part and who does not. When designing a survey, an important aspect, among others, is the planning of the fieldwork period and the amount of time available for data collection. The manner in which data is collected (interviewers coming at the door, by telephone or through a web survey) can also affect who participates and who does not. In chapter 3 two surveys are compared that differ in the degree to which they took into account the living situation of non-Western minorities. Adapting the survey to the living situation may, for instance, include offering translated questionnaires, using bilingual interviewers with a shared ethnic background or establishing shorter fieldwork periods because it is more common among non-Western minorities a) to have problems with understanding Dutch, b) especially among older members of the non-Western communities, to be functionally illiterate and c) to move house more frequently. Chapter 4 compares two surveys that differ in the manner of data collection. One survey only uses interviewers that come at the door, combined with tailor-made measures, such as the use of translated questionnaires and the use of bilingual interviewers with the same ethnic background as the intended respondents (the interviewer approach). In the other survey, the intended respondent was first asked to complete the questionnaire on the web. If the intended respondent did not take part through the web, they would later be called for a telephone interview. In case that approach still wasn't successful, an interviewer would come at the door and ask the intended respondent to participate in an in-person interview. This last survey used the same tailor-made measures and the fieldwork period was equally long (the combined approach).

Based on a wide scale of methods for analysis, we could draw the following conclusions. Surveys that take into account the living situation of non-Western minorities to a greater degree give a more representative image of the target group than a more standard approach. Such surveys are not so much focused on obtaining a higher response as such, but are more focused on ensuring that the response is more balanced across the different subgroups that we are interested in: everyone can participate. Using bilingual interviewers with a shared ethnic background results in an improvement of the representation of the non-Western minority population in The Netherlands: this helps obtain a better representation of the first generation of migrants. Finally, it turns out that, in the comparison between the combined and the interviewer approach, the first approach leads to higher response rates, but not to a more balanced representation of the target group, even when translated questionnaires and interviewers from the same ethnic group are used in the last phase.

In chapter 5, the topic of interest was whether the manner of administering a survey question affected the respondent's answers. The study concerned the combined approach with translated questionnaires and bilingual interviewers with a shared ethnic background, looking at the possible influence of the data collection method on the answers. It appears that there are systematic differences depending on the method of collection, web surveys delivering the most honest answers among respondents. We are confronted here with a dilemma: using bilingual interviewers with the same ethnic background that come to the door is necessary in order to get a representative image of ethnic minority populations in which linguistic problems and/or functional illiteracy are more common, but apparently this approach generates more socially desirable answers.

Chapter 6 focuses on the issue of the comparability of survey data. In this chapter, we first analyzed to what degree undesired methodological measurement effects, like those of the ethnic background or the gender of the interviewer, or that of the language in which the interview is conducted, are of influence on the answers of the respondents. The question in this context was whether these undesired measurement effects were affecting each ethnic group just as much or not at all, or else, whether this influence was different between groups. In the first case, the comparability of the data is not compromised, but in the second case, the comparison is biased.

The analyses showed that the comparison between groups was biased. The answers of certain ethnic groups were more heavily influenced by undesired methodological measurement effects than those of other groups. For instance, Turkish male respondents seemed to answer more traditionally when they were interviewed by a male interviewer of Turkish origin, while this effect was not observable in the case of Turkish women or in any of the other ethnic groups. As a consequence, Turkish men seemed to be, on average, comparatively more traditional, while this was partly a methodological artefact. The 7th and last chapter summarizes the important conclusions regarding the main research question of the thesis. This summary is followed by a discussion of the reasons why these results are relevant and for whom they may be of interest.

About the author

Joost Kappelhof was born on the 27th of August, 1974, in Hoorn, The Netherlands. He studied Statistics and Methodology at the Psychology department of the University of Amsterdam. He also completed courses on Statistics and Algebra at the Mathematics department of the University of Amsterdam. After graduation in 2001, he started to work for The Netherlands Institute for Social Research/SCP as a survey methodologist in December 2001. At The Netherlands Institute for Social Research/SCP he has been extensively involved in survey design and survey data quality issues in cross-national and cross-cultural survey research.

Reference List

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

Anderson, B. A., Silver, B. D., and Abramson, P. R. (1988). The effects of the race of the interviewer on race-related attitudes of black respondents in SRC/CPS national election studies. Public Opinion Quarterly, 52, 3, 289-324.

Andridge, R. R. and Little, R. J. (2011). Proxy pattern-mixture analysis for survey nonresponse. Journal of Official Statistics 27, 2, 153-180.

Arends-Tóth, J. and Van de Vijver, F. J. (2008). Family relationships among immigrants and majority members in the Netherlands: The role of acculturation. Applied Psychology, 57, 466-487.

Aspinall, P. J. (2002). Collective Terminology to Describe the Minority Ethnic Population. The Persistence of Confusion and Ambiguity in Usage. Sociology, 36, 4, 803-816. Doi: 10.1177/003803850203600401

Barnes, W. (2008). Improving Migrant Participation in the Abour Force Survey: A review of Existing Practices in European Union Member States. Survey Methodology Bulletin, 63, 25-38.

Baumgartner, H. and Steenkamp, J. B. E. (2001). Response styles in marketing research: A cross-national investigation. Journal of Marketing Research, 38, 2, 143-156.

Beatty, P. C. and Willis, G. B. (2007). Research synthesis: the practice of cognitive interviewing. Public Opinion Quarterly, 71, 2, 287-311.

Bentler, P. M. (1990). Comparative fit indexes in structural models. Psychological bulletin, 107, 2, 238-246.

Bethlehem, J. G. (1988). Reduction of nonresponse bias through regression estimation. Journal of Official Statistics, 4, 3, 251-260.

Bethlehem, J., Cobben, F., and Schouten, B. (2011). Handbook of nonresponse in household surveys. (Wiley Handbooks in Survey Methodology). Hoboken, New Jersey: John Wiley and Sons, Inc.

Bhopal, R. (2004). Glossary of terms relating to ethnicity and race: for reflection and debate. Journal of Epidemiol Community Health, 58, 441-445. Doi:10.1136/jech.2003.013466

Bia, M. and Mattei, A. (2008). A Stata package for the estimation of the dose-response function through adjustment for the generalized propensity score. The Stata Journal, 8, 354-373.

Biemer, P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. Journal of Official Statistics, 17, 2, 295-320.

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

Biemer, P. (2010). Total survey error: Design, implementation, and evaluation. Public opinion quarterly, 74, 5, 817-848.

Bijl, R. and Verweij, A. (2012). Measuring and monitoring immigrant integration in Europe. The Netherlands, The Hague: the Netherlands Institute for Social Research/SCP.

Billiet, J. and Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. Sociological Methods and Research, 36, 4, 542-562.

Billiet, J. and McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. Structural equation modeling: A Multidisciplinary Journal, 7, 4, 608-628.

Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. Journal of Human resources, 8, 4, 436-455.

Boehnke, K., Lietz, P., Schreier, M., and Wilhelm, A. (2011). Sampling: The selection of cases for culturally comparative psychological research. In D. Matsumoto and F. J. R. Van de Vijver (Eds.), Cross-cultural research methods in psychology (pp. 101-129). New York: Cambridge University Press.

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality, Journal of Public Health, 27, 281-291.

Braun, M. and Johnson, T. P. (2010). An illustrative review of techniques for detecting inequivalences. In J. A.Harkness, M. Braun, B. Edwards, T. Johnson, L. Lyberg, P. Mohler, B. Pennell, and T. Smith (Eds.), Survey methods in multinational, multiregional, and multicultural contexts (pp. 375-393). Wiley Online Library.

Brehm, J. (1993). The phantom respondents. Ann Arbor: University of Michigan Press.

Breslau, J., Kendler, K. S., Su, M., Gaxiola-Aguilar, S., and Kessler, R. C. (2005). Lifetime risk and persistence of psychiatric disorders across ethnic groups in the United States. Psychological medicine, 35, 3, 317-327.

Buelens, B. and Van den Brakel, J. (2011). Inference in surveys with sequential mixed-mode data collection. Discussion paper, 201121, The Hague/Heerlen: CBS/Statistics Netherlands.

Caetano, R., Clark, C. L., and Tam, T. (1998). Alcohol consumption among racial/ethnic minorities. Alcohol health and research world, 22, 4, 233-241.

Campbell, B. A. (1981). Race-of-interviewer effects among southern adolescents. Public Opinion Quarterly, 45,2, 231-244.

CBS-Statline. 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.

Chaturvedi, N. and McKeigue, P. M. (1994). Methods for epidemiological surveys of ethnic minority groups. Journal of epidemiology and community health, 48, 107-111. DOI:10.1136/jech.48.2.107

Chen, C., Lee, S. y., and Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. Psychological Science, 6, 3, 170-175.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., and Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. Sociological Methods and Research, 36, 4, 462-494.

Cheung, G. W. and Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. Journal of Cross-Cultural Psychology, 31, 2, 187-212.

Cotter, P. R., Cohen, J., and Coulter, P. B. (1982). Race-of-interviewer effects in telephone interviews. Public Opinion Quarterly, 46,2, 278-284.

Couper, M. P. (2005). Technology trends in survey data collection. Social Science Computer Review 23, 4, 486-501. DOI: http://dx.doi.org/ 10.1177/0894439305278972

Couper, M. P. (2011). The future modes of data collection. Public Opinion Quarterly, 75, 5, 889-908.

Couper, M. P., Tourangeau, R., Conrad, F. G., and Crawford, S. D. (2004). What They See Is What We Get Response Options for Web Surveys. Social Science Computer Review, 22, 1, 111-127.

Curtain, R., Presser, S., and Singer, S. (2000). The effect of Response Rate Changes on the Index of Consumer Sentiment, Public Opinion Quarterly, 64, 4, 413-428.

Dagevos, J. and Gijsberts, M. (2009). Social-culture positie. In M. Gijsberts and J. Dagevos (Eds.), Jaarrapport Integratie 2009 (pp. 226-253). [In Dutch: Sociocultural position]. Den Haag: SCP.

Dagevos, J., Gijsberts, M., Kappelhof, J. and Vervoort, M. (2007). Survey Integratie Minderheden 2006. Verantwoording van de opzet en de uitvoering van een survey onder Turken, Marokkanen, Surinamers, Antillianen en een autochtone vergelijkingsgroep [In Dutch: Survey on the integration of minorities 2006. Design and fieldwork]. Den Haag: SCP. http://www.scp.nl/Publicaties/Alle_publicaties/Publicaties_2007/Survey_integratie_minderheden_2006. (accessed June 2012)

Dagevos, J. and Schellingerhout, R. (2003). Sociaal-culturele integratie. Contacten, cultuur en orientatie op de eigen groep. In J. Dagevos, M. Gijsberts, and C. van Praag (Eds.), Rapportage minderheden (pp. 317-362). [In Dutch: Sociocultural 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).

Dagevos, J., Schellingerhout, R., and Vervoort, M. (2007). Sociaal-culturele integratie en religie. In J. Dagevos and M. Gijsberts (Eds.), Jaarrapport Integratie 2007 (pp. 163-191). [In Dutch: Sociocultural integration and religion] Den Haag: SCP.

Davidov, E., Schmidt, P., and Billiet, J. (2011). Cross-cultural analysis: Methods and applications. London, England: Routledge.

Davis, D. W. (1997). The direction of race of interviewer effects among African-Americans: Donning the black mask. American Journal of Political Science, 41, 1, 309-322.

Davis, R. E., Couper, M. P., Janz, N. K., Caldwell, C. H., and Resnicow, K. (2010). Interviewer effects in public health surveys. Health Education Research, 25, 1, 14-26.

Deding, M., Fridberg, T., and Jakobsen, V. (2008). Non-response in a survey among immigrants in Denmark. Survey Research Methods, 2, 3, 107-121.

Deding, M., Fridberg, T., and Jakobsen, V. (2013). Non-response among immigrants in Denmark. In J. Font and M. Mendez (Eds.), Surveying Ethnic Minorities and Immigrant populations (pp. 173-194). Amsterdam: Amsterdam University Press.

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

De Heer, W. and De Leeuw, E. D. (2001). Trends in Household Survey Nonresponse: A longitudinal and international comparison. In R. M. Groves, D. Dillman, J. L. Eltinge, and R. J. A. Little (Eds.). Survey Nonresponse (pp. 41-54). New York, US: Wiley-Blackwell.

De Leeuw, E. D. (1992). Data Quality in Mail, Telephone and Face to Face Surveys. Amsterdam: TT-Publikaties.

De Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. Journal of Official Statistics, 21, 2, 233-255.

De Leeuw, E. D., Dillman, D. A., and Hox, J. J. (2008). Mixed-Mode surveys: When and why. In E. D. de Leeuw, J. J. Hox, and D. A. Dillman (Eds.), International Handbook Of Survey Methodology (pp. 299-316). New York: Taylor and Francis Group.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39, 1, 1-38.

Dillman, D. A. (2000). Mail and internet surveys: The tailored design method. New York: John Wiley and Sons Inc.

Dillman, D. A. (2007). Mail and internet surveys: The tailored design method. (2nd ed.). Hoboken, New Jersey: John Wiley and Sons, Inc.

Dillman, D. A. and 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. et al. (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.

DiNardo, J. (2002). Propensity score reweighting and changes in wage distributions. Mimeo. http://www-personal.umich.edu/~jdinardo/bztalk5.pdf.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. Econometrics, 64, 5, 1001-1044.

Dourleijn, E. (2010). Survey Integratie Nieuwe Groepen 2009. Verantwoording van de opzet en uitvoering van een survey onder Afghaanse, Iraanse, Iraakse, Somalische, (kort verblijvende) Poolse en Chinese Nederlanders en een autochtone Nederlandse vergelijkingsgroep. [In Dutch: fieldwork report of a survey among Afghan, Iranian, Iraki, Somalian, (short stay) Polish and Dutch of Chinese descent and native Dutch] Den Haag: Sociaal en Cultureel Planbureau.

Dutwin, D. and Lopez, M. H. (2014). Considerations of survey error in surveys of Hispanics. Public opinion quarterly, 78, 2, 392-415.

Duque, I., Ballano, C., and Perez, C. (2013). The 2007 Spanish National Immigrant Survey (ENI): Sampling from the Padrón. In J. Font and M. Mendez (Eds.), Surveying Ethnic Minorities and Immigrant populations (pp. 69-84). Amsterdam: Amsterdam University Press.

Erens, B. (2013). Designing high-quality surveys of ethnic minority groups in the United Kingdom. In J. Font and M. Mendez (Eds.), Surveying Ethnic Minorities and Immigrant populations (pp. 45-68). Amsterdam: Amsterdam University Press.

EUROSTAT. (2000). Assessment of the quality in statistics. Eurostat/A4/Quality/00/ General/Standard report.

Fernandez, A. L. and Marcopulos, B. A. (2008). A comparison of normative data for the Trail Making Test from several countries: Equivalence of norms and considerations for interpretation. Scandinavian journal of psychology, 49, 3, 239-246.

Feskens, R. C. W. (2009). Difficult Groups in Survey Research and the Development of Tailor-made Approach Strategies. The Hague: Statistics Netherlands/University Utrecht. Retrieved from http://www.cbs.nl/NR/rdonlyres/8F317AA9-1074-4BF7-84EB-015D013DBB80/0/2009x11feskenspub.pdf

Feskens, R., Hox, J., Lensvelt-Mulders, G., and Schmeets, H. (2006). Collecting data among ethnic minorities in an international perspective. Field Methods, 18, 3, 284-304.

Feskens, R., Hox, J., Lensvelt-Mulders, G. J. L. M., and Schmeets, H. (2007). Nonresponse among ethnic minorities: a multivariate analysis. Journal of Official Statistics, 23, 3, 387 -408.

Feskens, R. C. W., Kappelhof, J., Dagevos, J., and Stoop, I. A. L. (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

Finkel, S. E., Guterbock, T. M., and Borg, M. J. (1991). Race-of-Interviewer Effects in a Preelection Poll Virginia 1989. Public Opinion Quarterly, 55, 3, 313-330.

Font, J. and Mendez, M. (2013a). Introduction: The methodological challenges of surveying populations of immigrant origin. In J. Font and M. Mendez (Eds.), Surveying Ethnic Minorities and Immigrant populations (pp. 11-44). Amsterdam: Amsterdam University Press.

Galesic, M. and Bosnjak, M. (2009). Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey. Public Opinion Quarterly, 73, 2, 349-360. Doi: 10.1093/poq/nfp031

Gijsberts, M. and Iedema, J. (2011). Opleidingsniveau van niet-schoolgaanden en leerprestaties in het basisonderwijs. In M. Gijsberts, W. Huijnk, and J. Dagevos (Eds.), Jaarrapport integratie 2011 (pp. 76-99). [In Dutch: educational attainment of persons not in school and educational achievements in primary school]. Den Haag: SCP. Available at: http://www.scp.nl/Publicaties/Alle_publicaties/Publicaties_2012/Jaarrapport_integratie_2011

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention Science, 8, 3, 206-213.

Groeneveld, S. and Weijers-Martens, Y. (2003). Minderheden in beeld: SPVA-02. [In Dutch: The focus on non-Western ethnic minorities: SPVA-02]. Rotterdam: Instituut voor Sociologisch-Economisch Onderzoek (ISEO).

Groves, R. M. (1989). Survey costs and survey errors. Hoboken, New Jersey: John Wiley and Sons, Inc.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. Public Opinion Quarterly, 70, 5, 646-675. Doi: 10.1093/poq/nfl033

Groves, R. M. and Couper, M. P. (1998). Nonresponse in household interview surveys. Hoboken, New Jersey: John Wiley and Sons, Inc.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). Survey methodology. Hoboken, New Jersey: John Wiley and Sons, Inc.

Groves, R. M. and Heeringa, S. G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. Journal of the Royal Statistical Society: Series A (Statistics in Society), 169, 3, 439-457.

Groves, R. M. and Lyberg, L. (2010). Total survey error: Past, present, and future. Public opinion quarterly, 74, 5, 849-879.

Groves, R. M. and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: a meta-analysis. Public Opinion Quarterly, 72, 2, 167-189.

Hatchett, S. and Schuman, H. (1975). White respondents and race of interviewer effects. Public Opinion Quarterly, 39, 4, 523-528.

Harkness, J. A. (2007). Improving the comparability of translations. In R. Jowell, C. Roberts, R. Fitzgerald, and G. Eva (Eds.). Measuring attitudes cross-nationally: Lessons from the European Social Survey (pp. 79-95). London: Sage publication Ltd.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T., Lijberg, L., Mohler, P. et al. (2010). Survey methods in multinational, multiregional, and multicultural contexts. Hoboken, New Jersey: John Wiley and Sons.

Harkness, J. A., Pennell, B. E., and Schoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin., J. Martin, and E. Singer (Eds.), Methods for testing and evaluating survey questionnaires (pp. 453-473). Hoboken, New Jersey: John Wiley and Sons, Inc.

Harkness, J. A. and Schoua-Glusberg, A. (1998). Questionnaires in translation. In J. A. Harkness (Ed.), ZUMA-Nachrichten Spezial No. 3. Cross-Cultural Survey Equivalence (pp. 87-127). Mannheim: ZUMA.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T., Lyberg, L., Mohler, P. et al. (2010). Survey methods in multinational, multiregional, and multicultural contexts. Hoboken, New Jersey: John Wiley and Sons, Inc.

Harkness, J. A., Schoebi, N., Joye, D., Mohler, P., Faass, T., and Behr, D. (2008). Oral translation in telephone surveys. In J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. de Leeuw, L. Japec, and P. J. Lavrakas (Eds.), Advances in telephone survey methodology (pp. 231-249). Hoboken, New Jersey: John Wiley & Sons.

Harkness, J. A., Van de Vijver, F. J. R., and Mohler, P. (2003). Cross-cultural survey methods. New York: John Wiley and Sons

He, J.,& Van de Vijver, F. (2012). Bias and Equivalence in Cross-Cultural Research. Online Readings in Psychology and Culture, 2, 2. http://dx.doi.org/10.9707/2307-0919.1111

He, J. and Van de Vijver, F. J. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. Personality and Individual Differences, 55, 7, 794-800.

Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: an experimental investigation of data quality and social desirability effects. International Journal of Public Opinion Research, 21, 1, 111-121.

Herníndez-Quevedo, C. and Jiminez-Rubio, D. (2009). A comparison of the health status and health care utilization patterns between foreigners and the national population in Spain: new evidence from the Spanish National Health Survey. Social Science and Medicine, 69, 3, 370-378.

Hilhorst, M. (2007). Survey integratie minderheden (SIM) 2006. [In Dutch: Survey on the Integration of Minorities: fieldwork report SIM2006] Bureau Veldkamp, Amsterdam.

Hox, J. J. and De Leeuw, E. D. (1994). Applying multilevel modelling to meta-analysis. Quality and Quantity, 28, 4, 329-344.

Hox, J. J., De Leeuw, E. D., and Brinkhuis, M. J. S. (2010). Analysis models for comparative surveys. In J. A. Harkness, M. Braun, B. Edwards, T. Johnson, L. Lyberg, P. Mohler, B. E. Pennell, and T. W. Smith (Eds.), Survey Methods in multinational, multiregional, and multicultural contexts (pp. 395-418). Hoboken, New Jersey: John Wiley & Sons.

Hu, L. and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling: A Multidisciplinary Journal, 6,1, 1-55.

Huang, I., Frangakis, C., Dominici, F., Diette, G. B., and Wu, A. W. (2005). Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. Health services research, 40, 1, 253-278.

Hui, C. H. and Triandis, H. C. (1983). Multistrategy Approach to Cross-Cultural Research The Case of Locus of Control. Journal of Cross-Cultural Psychology, 14, 1, 65-83.

Hui, C. H. and Triandis, H. C. (1985). Measurement in Cross-Cultural Psychology A Review and Comparison of Strategies. Journal of Cross-Cultural Psychology, 16, 2, 131-152.

Hui, C. H. and Triandis, H. C. (1989). Effects of culture and response format on extreme response style. Journal of Cross-Cultural Psychology, 20,3, 296-309.

Huijnk, W. and Dagevos, J. (2012). Dichter bij elkaar? De sociaal-culturele positie van niet-westerse migranten in Nederland. [In Dutch: Closer together? The sociocultural position of non-western migrants in the Netherlands]. SCP-2012-33. Den Haag: SCP.

Huijnk, W., Gijsberts, M. and Dagevos, J. (2014). Jaarrapport integratie 2013. [In Dutch: Year report Integration 2013]. SCP-2014-02. Den Haag: SCP.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. Biometrika, 87, 3, 706-710.

Jackle, A., Roberts, C., and Lynn, P. (2010). Assessing the effect of data collection mode on measurement. International Statistical Review, 78, 1, 3-20.

Jacobs, D., Swyngedouw, M., Hanquinet, L., Vandezande, V., Andersson, R., Beja Horta, A. P., Berger, M., Diani, M., Gonzalez Ferrer, A., Giugni, M., Morariu, M., Pilati, K., and Statham, P. (2009). The challenge of measuring immigrant origin and immigration-related ethnicity in Europe. International journal of Migration and Integration, 10, 1, 67-88. DOI 10.1007/s12134-009-0091-2

Jann, B. (2008). The Blinder-Oaxaca decomposition for linear regression models. The Stata Journal, 8, 453-479.

Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. ZUMA-Nachrichten spezial, 3, 1-40.

Johnson, T. P., Kulesa, P., Cho, Y. I., and Shavitt, S. (2005). The relation between culture and response styles evidence from 19 countries. Journal of Cross-Cultural Psychology, 36, 2, 264-277.

Johnson, T. P. and Van de Vijver, F. J. (2003). Social desirability in cross-cultural research. In : J. A. Harkness, F. J. van de Vijver, and P. Mohler (Eds.), Cross-cultural survey methods (pp. 193-202). New York: Wiley.

Joreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. Psychometrika, 36,2, 109-133.

Kalton, G. and Flores-Cervants, I. (2003). Weighting methods. Journal of Official Statistics, 19, 2, 81-97.

Kankaras, M. and Moors, G. (2009). Measurement equivalence in solidarity attitudes in Europe insights from a multiple-group latent-class factor approach. International Sociology, 24, 4, 557-579.

Kappelhof, J. W. S. (2010). Op maat gemaakt? Een evaluatie van enkele responsverbeterende maatregelen onder Nederlanders van niet-westerse afkomst. [In Dutch: An evaluation of several response-enhancing measures among Dutch of non-Western origin] SCP-special 53. SCP, Den Haag.

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

Kemper, F. (1998). Gezocht: Marokkanen. Methodische problemen bij het verwerven en interviewen van allochtone respondenten. [In Dutch: Wanted: Moroccans. Methodological problems with recruitment and interviewing nonnative respondents]. Migrantenstudies, 1, 43-57.

Kolenikov, S., and Kennedy, C. (2014). Evaluating Three Approaches to Statistically Adjust for Mode Effects. Journal of Survey Statistitics and Methodology, 2 2, 126-158. Doi: 10.1093/jssam/smu004

Korte, K. and Dagevos, J. (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 Ethnic Minorities 2011. Report on the design and fieldwork of a survey conducted among Dutch of Turkish, Moroccan, Surinamese and Antilean descent and a native Dutch control sample] Den Haag: SCP.

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

Kreuter, F., and Olson, K. (2013). Paradata for Nonresponse Error Investigation. In F. Kreuter (Ed.), Improving Surveys with paradata. Analytic Uses of Process Information (pp. 13-42). Wiley Series in Survey methodology. Hoboken, New Jersey: John Wiley & Sons.

Kreuter, F., Olson, K. M., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M., and Raghunathan, T. E. (2010). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-Response: Examples from Multiple Surveys. Sociology Department, Faculty Publications. Paper 137.Retrieved from http://digitalcommons.unl.edu/sociologyfacpub/137

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

Kuha, J. (2004). AIC and BIC comparisons of assumptions and performance. Sociological Methods and Research, 33, 2, 188-229.

Laganá, F., Elcheroth, G., Penic, S., Kleiner, B., and Fasel, N. (2013). National minorities and their representation in social surveys: which practices make a difference? Quality and Quantity, 47, 3, 1287-1314.

Lee, R. M. (1993). Doing research on sensitive topics. London, UK: Sage.

Leung, K., Lau, S., and Lam, W. L. (1998). Parenting styles and academic achievement: A cross-cultural study. Merrill-Palmer Quarterly (1982-), 44, 2, 157-172.

Liang, J., Asano, H., Bollen, K. A., Kahana, E. F., and Maeda, D. (1987). Cross-cultural comparability of the Philadelphia Geriatric Center Morale Scale: An American-Japanese comparison. Journal of Gerontology, 42, 1, 37-43.

Lipps, O. and Kissau, K. (2012). Nonresponse in an individual register sample telephone survey in Lucerne/Switzerland. In M. Hader, S. Hader, and M. Kuhne (Eds.), Telephone Surveys in Europe: Research and practice (Pp. 187-208). Berlin: Springer.

Lipps, O., Laganá, F., Pollien, A., and Gianettoni, L. (2013). Under-representation of foreign minorities in cross-sectional and longitudinal surveys in Switzerland. In J. Font and M. Mendez (Eds.), Surveying Ethnic Minorities and Immigrant populations (pp. 241-270). Amsterdam: Amsterdam University Press.

Little, R. J. and Rubin, D. B. (2002). Statistical analysis with missing data. Hoboken, New Jersey: John Wiley and Sons, Inc.

Lozar Manfreda, K., 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

Lubke, G. H., Dolan, C. V., Kelderman, H., and Mellenbergh, G. J. (2003). On the relationship between sources of within-and between-group differences and measurement invariance in the common factor model. Intelligence, 31, 6, 543-566.

Lubke, G. H. and Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. Structural equation modeling, 11, 4, 514-534.

Lubke, G. H. and Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. Psychological methods, 10, 1, 21-39.

Lynn, P. (2003). Developing quality standards for cross-national survey research: five approaches. International Journal of Social Research Methodology, 6, 4, 323-336.

Lynn, P., Sturgis, P., Clarke, P., and Martin, J. (2001). The effects of extended interviewer effort on non-response bias. In R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. Little (Eds.). Survey Nonresponse (pp. 135-148). New York, US: Wiley-Blackwell.

Maitland, A., Casas-Cordero, C., and Kreuter, F. (2009). An evaluation of nonresponse bias using paradata from a health survey. 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

Marin, G., Gamba, R. J., and Marin, B. V. (1992). Extreme Response Style and Acquiescence among Hispanics The Role of Acculturation and Education. Journal of Cross-Cultural Psychology, 23, 4, 498-509.

Martens, E. P. (1999). Minderheden in beeld: SPVA-98. [In Dutch: The focus on non-Western ethnic minorities: SPVA-98]. Rotterdam: NIWI.

Mateos, P., Webber, R., and Longley, P. (2007). The Cultural, Ethnic and Linguistic Classification of Populations and Neighbourhoods using Personal Names. CASA Working Paper 116. University College London.

McManus, S., Erens, B., and Bajekal, M. (2006). Conducting surveys among ethnic minority groups in Britain. In J. Y. Nazroo (Ed.), Health and social research in multiethnic societies (pp. 116-148). London: Routledge.

Mendez, M., Ferreras, M., and Cuesta, M. (2013). Immigration and general population surveys in Spain: The CIS surveys. In J. Font and M. Mendez (Eds.), Surveying Ethnic Minorities and Immigrant populations (pp. 195-218). Amsterdam: Amsterdam University Press.

Mellenbergh, G. J. (1989). Item bias and item response theory. International Journal of Educational Research, 13, 2, 127-143.

Mendez, M. and Font, J. (2013). Surveying immigrant populations: Methodological strategies, good practices and open questions. In J. Font and M. Mendez (Eds.), Surveying Ethnic Minorities and Immigrant populations (pp. 147-172). Amsterdam: Amsterdam University Press.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58, 4, 525-543.

Meredith, W. and Teresi, J. A. (2006). An essay on measurement and factorial invariance. Medical care, 44, 11, S69-S77.

Moorman, P. G., Newman, B., Millikan, R. C., Tse, C. K. J., and Sandler, D. P. (1999). Participation rates in a case-control study: The impact of age, race, and race of interviewer. Annals of epidemiology, 9, 3, 188-195.

Morales, L. and Ros, V. (2013). Comparing the response rates of autochthonous and migrant populations in nominal sampling surveys: The LOCALMULTIDEM study in Madrid. In J. Font and M. Mendez (Eds.), Surveying Ethnic Minorities and Immigrant populations (pp. 147-172). Amsterdam: Amsterdam University Press.

Morren, M., Gelissen, J. P., and Vermunt, J. K. (2011). Dealing with extreme response style in cross-cultural research: a restricted latent class factor analysis approach. Sociological Methodology, 41, 1, 13-47.

Morren, M., Gelissen, J., and Vermunt, J. K. (2012a). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 8, 4, 159 - 170.

Morren, M., Gelissen, J. P., and Vermunt, J. K. (2012b). Response Strategies and Response Styles in Cross-Cultural Surveys. Cross-Cultural Research, 46, 3, 255-279.

Muthén, L. K. and Muthén, B. O. (2011). Mplus User's Guide. Sixth Edition. [Computer software]. Los Angeles, CA.

Myrberg, G. (2013). Surveying migrants and migrant associations in Stockholm. In J. Font and M. Mendez (Eds.), Surveying Ethnic Minorities and Immigrant populations (pp. 131-146). Amsterdam: Amsterdam University Press.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. International economic review, 14, 3, 693-709.

OECD. (2011). Quality dimensions, core values for OECD statistics and procedures for planning and evaluating statistical activities. OECD.

Okazaki, S. and Sue, S. (1995). Methodological issues in assessment research with ethnic minorities. Psychological Assessment, 7, 3, 367-375.

Olson, K. (2013). Paradata for Nonresponse Adjustment. Annals of the American Academy of Political and Social Sciences, 645, January 2013, 142-170.

Pierzchala, M. (2006). Disparate modes and their effect on instrument design. In 10th International Blaise Users Conference. Papendal-Arnhem, 199-209.

Poortinga, Y. H. and Van de Vijver, F. J. (1987). Explaining Cross-Cultural Differences Bias Analysis and Beyond. Journal of Cross-Cultural Psychology, 18, 3, 259-282.

Reep, C. (2003). Moeilijk Waarneembare Groepen. Een inventarisatie. [In Dutch: Hard to survey populations: An inventory]. Rep. No. H1568-03-s00. Voorburg/Heerlen: CBS.

Rinken, S. (2013). Enhancing representativeness in highly dynamic settings: Lessons from the NEPIA survey. In J. Font and M. Mendez (Eds.), Surveying Ethnic Minorities and Immigrant populations (pp. 85-110). Amsterdam: Amsterdam University Press.

Rhodes, P. J. (1994). Race-of-interviewer effects: a brief comment. Sociology, 28, 547-558.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 1, 41-55.

Ross, C. E. and Mirowsky, J. (1984). Socially-desirable response and acquiescence in a cross-cultural survey of mental health. Journal of Health and Social Behavior, 25, 2, 189-197.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. Hoboken, New Jersey: John Wiley and Sons, Inc.

Saris, W. E. and Gallhofer, I. N. (2014). Design, evaluation and analysis of questionnaires for survey research. (2nd ed). Hoboken, New Jersey: John Wiley and Sons, Inc.

Särndal, C. (2011). Three Factors to Signal Non Response Bias With Applications to Categorical Auxiliary Variables. International Statistical Review 79, 2, 233-254. DOI: 10.1111/j.1751-5823.2011.00142.x

Särndal, C. and Lundström, S. (2005). Estimation in Surveys with Nonresponse. Chichester, England: John Wiley and Sons.

Särndal, C. and Lundström, S. (2010). Design for estimation: identifying auxiliary vectors to reduce nonresponse bias. Survey Methodology, 36, 2, 131-144.

Schaeffer, N. C., (1980). Evaluating race of interviewer effect in a national survey. Sociological Methods and Research, 8, 4, 400-413.

Schmeets, H. (2005). De leefsituatie van allochtonen.[In Dutch: The living conditions of nonnatives]. In H. Schmeets and R. van der Bie (Eds.), Enqueteonderzoek onder allochtonen. Problemen en oplossingen (pp. 169-176). Voorburg/Heerlen: CBS.

Schmeets, H. and Van der Bie, R. (2005). Enqueteonderzoek onder allochtonen. Problemen en oplossingen. [In Dutch: survey research among minorities. Problems and solutions]. Voorburg/Heerlen: CBS.

Schnell, R., Gramlich, T., Bachteler, T., Reiher, J., Trappmann, M., Smid, M., and Becher, I. (2013). Ein neues Verfahren für namensbasierte Zufallsstichproben von Migranten. [In German: A new Name-Based Sampling Method for Migrants using n-grams] MDA – Methoden – Daten – Analysen, 7, 1, 5-33.

Scholten, P. (2011). Framing Immigrant Integration: Dutch Research-Policy Dialogues in Comparative Perspective. Amsterdam: Amsterdam University Press.

Schothorst, Y. (2002). Onderzoek onder allochtonen: wat kan, wat moet en wat kan? In H. Houtkoop and J. Veenman (Eds.), Interviewen in de multiculturele samenleving: problemen en oplossingen (pp. 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 Cobben, F. (2007). R-indexes for the comparison of different fieldwork strategies and data collection modes. Rep. No. Discussion Paper 07002. Voorburg/Heerlen:CBS. Retrieved from: http://www.risq-project.eu/papers/schouten-cobben-2007-a.pdf (last accessed October 2013).

Schouten, B. and Cobben, F. (2008). An empirical validation of R-indicators. Rep. No. Discussion Paper 08006. Voorburg, the Netherlands: Statistics Netherlands. Retrieved from http://www.risq-project.eu/papers/cobben-schouten-2008-a.pdf (last accessed October 2013).

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators of Representativity of Survey Nonresponse. Survey Methodology, 35, 1, 101-113.

Schouten, B., Bethlehem, J., Beullens, K., Kleven, Ø., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N., and Skinner, C. (2013). Evaluating, Comparing, Monitoring, and Improving Representativity of Survey Response Through R-Indicators and Partial R-Indicators. International Statistical Review, 80, 382-399.

Schouten, B., Luiten, A., Loosveldt, G., Beullens, K., and Kleven, Ø. (2010). Monitoring and changing data collection through R-indicators and partial R-indicators. Rep. No. RISQ deliverable 10.

Schouten, B., Shlomo, N., and Skinner, C. (2011). Indicators for monitoring and improving representativity of response. Journal of Official Statistics, 27, 231-253.

Schuman, H. and Converse, J. M. (1971). The effects of black and white interviewers on black responses in 1968. Public Opinion Quarterly, 35, 1, 44-68.

Shlomo, N., Skinner, C., Schouten, B., Carolina, T., and Morren, M. (2009). Partial indicators for representative response. Rep. No. RISQ deliverable 4. version 2.

Simon, P. (2007). "Ethnic" statistics and data protection in the Council of Europe countries. Study Report. Strasbourg: Conseil de l'Europe.

Singer, E. (2001). The use of incentives to reduce nonresponse in household surveys. In R. M. Groves, D. Dillman, J. L. Eltinge, and R. J. Little (Eds.), Survey nonresponse (pp. 163-177). New York: John Wiley and Sons.

Singer, E., Van Hoewyk, J., Gebler, N., Raghunathan, T., and McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. Journal of Official Statistics, 15, 2, 217-230.

Singer, E., Van Hoewyk, J., and Maher, M. P. (2000). Experiments with incentives in telephone surveys. Public Opinion Quarterly, 64, 2, 171-188.

Sinibaldi, J., Durrant, G. B., and Kreuter, F. (2013). Evaluating the Measurement Error of Interviewer Observed Paradata. Public Opinion Quarterly, 77, S1, 173-193

Smith, T. W. (2013). An evaluation of Spanish questions on the 2006 and 2008 US General Social Surveys. In J. Font and M. Mendez (Eds.), Surveying Ethnic Minorities and Immigrant populations (pp. 219-240). Amsterdam: Amsterdam University Press.

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.

Steiger, J. H. (1989). EzPATH: Causal modeling. Evanston, IL: SYSTAT.

Stoop, I. A. L. (2005). The hunt for the last respondent: Nonresponse in sample surveys. The Netherlands, The Hague: the Netherlands institute for Social Research/SCP.

Stoop, I.A.L. (2014). Representing the populations: what general social surveys can learn from surveys among specific groups. In R. Tourangeau, B. Edwars, T. P. Johnson, K. M. Wolter, and N. Bates (Eds.), Hard-to-Survey Populations (pp. 225-244). Cambridge, U.K: Cambridge University Press.

Stoop, I.A.L., Billiet, J., Koch, A., and Fitzgerald, R. (2010). Improving survey response: Lessons learned from the European Social Survey. Chichester, UK: John Wiley and Sons.

Sudman, S. and Bradburn, N. M. (1974). Response effects in surveys: A review and synthesis. Aldine Publishing Company Chicago, Ill.

Survey Research Center (2010). Guidelines for Best Practice in Cross-Cultural Surveys. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved 01-02-2014, from http://www.ccsg.isr.umich.edu/.

Suzer-Gurtekin, Z. T., Heeringa, S. G., and Valliant, R. (2012). Investigating the Bias of Alternative Statistical Inference Methods in Sequential Mixed-Mode Surveys. Proceedings of the Survey Research Methods Section. ASA, 4711-4725.

Thomas, M. (2008). Improving Immigrant participation in the labour force survey: Non-response and Attitudes of Non-English speaking Migrants to participation. Survey Methodology Bulletin, 63, 39-51.

Thornberry, O. T. and Massey, J. T. (1988). Trends in United States telephone coverage across time and subgroups. New York: John Wiley and Sons, Inc.

Tourangeau, R. (2013). Confronting the challenges of household surveys by mixing modes. Keynote address at the 2013 Federal Committee on Statistical Methodology Research Conference, Nov 4, 2013. Retrieved 11-18-2014, from http://www.copafs.org/UserFiles/file/fcsm/Tourangeau2013FCSMResearchKeynote.pdf

Tourangeau, R., Conrad, F. R. and Couper, M. P. (2013). The science of web surveys. New York: Oxford University Press.

Tourangeau, R., Rips, L. J., and Rasinski, K. A. (2000). The psychology of survey response. Cambridge University Press.

Tourangeau, R., and Yan, T. (2007). Sensitive questions in surveys. Psychological bulletin, 133, 5, 859-883.

Tucker, L. R. and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1, 1-10.

Vannieuwenhuyze, J. and Molenberghs, G. (2010). A SAS macro to disentangle mode effects on proportions and the mean of a categorical variable in an extended mixed-mode dataset. KU Leuven: Ceso.

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

Vannieuwenhuyze, J., Loosveldt, G., and Molenberghs, G. (2012). A method to evaluate mode effects on the mean and variance of a continuous variable in mixed-mode surveys. International Statistical Review, 80, 2, 306-322.

Van de Vijver, F. J. R. (2003). Bias and equivalence: Cross-cultural perspectives. In J. A. Harkness, F. J. R. Van de Vijver, and P. Mohler. (Eds.), Cross-cultural survey methods (pp. 143-155). Hoboken, New Jersey: John Wiley & Sons.

Van de Vijver, F. J.R. (2011). Capturing bias in structural equation modeling. In E. Davidov, P. Schmidt, and J. Billiet (Eds.), Cross-cultural analysis. Methods and applications (pp. 3-34). London, England: Routledge.

Van de Vijver, F. J. R., and Leung, K. (1997). Methods and data analysis of comparative research. In J.W. Berry, Y. H. Poortinga, and J. Pandey (Eds.), Handbook of cross-cultural psychology, Vol. 1: Theory and method (2nd ed.) (pp. 257-300). Needham Heights, MA, US: Allyn & Bacon, xxv.

Van de Vijver, F. J. R., and Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. Revue Europeenne de Psychologie Appliquee/European Review of Applied Psychology, 54, 2, 119-135.

Van der Zouwen, J. (2006). De interviewer, hulp of hindernis? In A. E. Bronner, P. Dekker, E. D. de Leeuw, L. J. Paas, K. de Ruyter, A. Smidts, and J. E. Wieringa (Eds.), Ontwikkelingen in het Marktonderzoek, Jaarboek 2006 (pp. 63-76). [In Dutch: The Interviewer: help or impediment?] Haarlem: Spaarenhout.

Vandenberg, R. J. and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational research methods, 3, 1, 4-70.

Van Heelsum, A. J. (1997). De etnisch-culturele positie van de tweede generatie Surinamers. Doctoral Dissertation. Amsterdam: Free University. http://hdl.handle.net/1871/13062

Van Heelsum, A.J. (2013). The influence of interviewers' ethnic background in a survey among Surinamese in the Netherlands. In J. Font and M. Mendez (Eds.), Surveying Ethnic Minorities and Immigrant populations (pp. 111-130). Amsterdam: Amsterdam University Press.

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: the Netherlands Institute for Social Research/SCP. (SCP- 2007/24).

Van't Land, H. (2000). Similar Questions: Different Meanings. Differences in the Meaning of Constructs for Dutch and Moroccan Respondents; Effects of the Ethnicity of the Interviewer and Language of the Interview among First and Second Generation Moroccan Respondents. Amsterdam: Vrije Universiteit.

Veenman, J. (2002). Interviewen in multicultureel Nederland. In H. Houtkoop and J. Veenman (Eds.), Interviewen in de multiculturele samenleving: problemen en oplossingen (pp. 1-19) [In Dutch; Interviewing in multicultural the Netherlands]. Assen: Koninklijke Van Gorcum.

Voogt, R. J. J., and Saris, W. E. (2005). Mixed-mode designs: Finding the balance between nonresponse bias and mode effects. Journal of Official Statistics, 21, 3, 367-387.

Wagner, J. (2008). Adaptive Survey Design to Reduce Nonresponse Bias. PhD thesis, University of Michigan.

Wagner, J. (2010). The fraction of missing information as a tool for monitoring the quality of survey data. Public Opinion Quarterly 74, 2, 223-243.

Weltevree, A. M., Boom, D. J., Rezai, S., Zuijderwijk, L., and Engbersen, G. (2009). Arbeidsmigranten uit Midden- en Oost-Europa. een profielschets van recente arbeidsmigranten uit MOE-landen. [In Dutch: Labour migrants from mid and eastern Europe. A profile of recent labour migrant from MOE-countries]. Rotterdam: Risbo.

Wicherts, J. M. (2007). Group Differences in Intelligence Test Performance. Universiteit van Amsterdam, Amsterdam.

Williams Jr, J. A. (1964). Interviewer-respondent interaction: A study of bias in the information interview. Sociometry, 27, 338-352.

WRR. (1989). Allochtonenbeleid. [In Dutch: minority policy]. Rapporten aan de Regering 36. Den Haag: SDU Uitgeverij.

Wen, S. W., Goel, V., and Williams, J. I. (1996). Utilization of health care services by immigrants and other ethnic/cultural groups in Ontario. Ethnicity and health, 1, 1, 99-109.