methods, data, analyses | Vol. 8(1), 2014, pp. 79-118 DOI: 10.12758/mda.2014.004

# The Impact of Method Bias on the Cross-Cultural Comparability in Face-to-Face Surveys Among Ethnic Minorities

*Joost W. S. Kappelhof*
*The Netherlands Institute for Social Research/SCP*

## Abstract

This article 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 socio-demographic 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.

*Keywords*: methods bias, non-Western ethnic minorities, cross-cultural comparative survey research; incomparability of samples, interviewer effects, multi group Mimic, socio-cultural integration

# Introduction

In general population surveys, non-Western minorities – or ethnic minorities as they are sometimes referred to – tend to be underrepresented (Feskens, 2009; Groves & Couper, 1998; Schmeets & Van der Bie, 2005). Ethnic minorities are difficult to survey mainly because of cultural differences, language barriers, socio-demographic 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 (Groeneveld & Weijers-Martens, 2003; Kappelhof, accepted; 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 questions, especially with respect to more sensitive questions (Tourangeau & 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 & Davidov, 2008; Cheung & Rensvold, 2000), social desirability (Johnson & Van de Vijver, 2003) or extreme response styles (Morren et al., 2012a; Morren et al., 2011; 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

*Direct correspondence to*
Joost W.S. Kappelhof, The Netherlands Institute for Social Research/SCP, The Hague, P.O. Box 16164, The Netherlands. E-mail: j.kappelhof@scp.nl

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 & Leung, 1997). This may cause problems, especially if one is interested in attempting to isolate ‘true’ cultural differences from differences in socio-demographic 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 socio-demographic point of view.

In the present study we investigate how these different factors affect the cross-cultural comparison of two socio-cultural integration constructs – attitudes towards *Gender Roles* and attitudes on *Familiy 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 & Yan, 2007). Socio-cultural 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 article sets out to investigate:

1. how interviewer effects influence the cross-cultural comparison of attitudes on Gender Roles and Familiy 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 *Familiy 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 *Familiy 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 *Familiy 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 *Familiy Ties* between non-Western groups in the Netherlands;

3. to what degree the observed differences in attitudes on *Gender Roles* and *Familiy Ties* between non-Western groups can be attributed to differences in socio-demographic 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 article 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.

# 1 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 & Van de Vijver, 2012; Poortinga & Van de Vijver, 1987; Van de Vijver & Leung, 1997; Van de Vijver & 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 & 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 & Lance, 2000). Three types of measurement equivalence are commonly distinguished for the measurement model (van de Vijver & Leung, 1997; van de Vijver & Tanzer, 2004).[1]

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, de Leeuw & Brinkhuis, 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 & Teresi, 2006; Vandenberg & 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 & Moors, 2009; Van de Vijver, 2003; Van de Vijver, 2011; Van de Vijver & Leung, 1997; Van de Vijver & 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.

[1] 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).

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 socio-demographic 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 & Steenkamp, 2001; Billiet & Davidov, 2008; Billiet & McClendon, 2000; Chen et al., 1995; Cheung & Rensvold, 2000; He & Van de Vijver, 2013; Hui & Triandis, 1989; Johnson et al., 2005; Marin et al., 1992; Morren et al., 2011; Morren et al., 2012a; Morren et al., 2012b; Ross & 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 socio-economic position and socio-cultural 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 & 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 & Bradburn, 1974; Schuman & Converse, 1971; Van Heelsum, 1997; Van't Land, 2000).Furthermore, a meta-analysis on sensitive questions in surveys by Tourangeau & 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 & Van de Vijver, 2012; Kankaras & Moors, 2009). Several studies have analyzed the impact of different socio-demographic sample composition of the compared cultural groups on the observed cross-cultural differences (Arends-Tóth & Van de Vijver, 2008; Fernandez & 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 & 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).

# 2 Data & Methods

## 2.1 Data

The data used in this article come from the Dutch Survey on the Integration of Minorities (SIM) that sets out to measure the socio-economic position of non-Western minorities as well as their socio-cultural 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

*Table 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 | 1565 |
| Moroccan | 48.0 | 829 | 1740 |
| Surinamese | 41.0 | 780 | 1930 |
| Antillean (incl. Aruban). | 44.2 | 863 | 1974 |

strata; Dutch of Turkish, Moroccan, Surinamese, and Antillean[2] descent and the remainder of the population (mostly native Dutch) living in the Netherlands, in the age of 15 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[3] 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 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 article.

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 1. Table 1 also includes, the gross sample and the sample size of each of the four response samples (i.e., the sample of the respondents).

In this article the SIM2011 response data file will be used. The response data file contains respondents' answers to survey questions, but also socio-demographic information on the respondent, socio-demographic information on the interviewer and interviewer observations (Table 2). Six survey questions measuring socio-cultural integration will be used in this analysis. These questions or a slightly larger

[2] Including Aruba

[3] or Aruba

set of questions have been used to measure socio-cultural integration of non-Western ethnic minorities in the Netherlands for over a decade (Arends-Tóth & Van de Vijver, 2008; Dagevos & Gijsberts, 2009; Dagevos & 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 Familiy 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.

## 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 *Familiy 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

*Table 2*: SIM2011 data used in the analysis

*Questions on socio-cultural 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).

*Socio-demographic 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 a Children (yes; no)
- Has a Partner (yes; no)
- Weight variable (design weight plus nonresponse adjustment)

*Socio-demographic 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.

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 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 *Familiy 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 *Familiy Ties*. Secondly, we expect that the shared ethnic background elicits more traditional views toward *Gender Roles* and *Familiy 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 *Familiy Ties*. If the interview language is Dutch, this will lead to less traditional views towards *Gender Roles* and *Familiy 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 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 socio-cultural 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 *Familiy 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 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 *Familiy 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 *Familiy 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 socio-demographic composition.

## 2.2 Methods

A variety of different modeling and analysis techniques have been used to detect equivalence of measures in cross-cultural research. See Braun & 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 socio-cultural 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 2). The latent variable *Family Ties* is measured by THUISHUW, VERTRFAMA and KIBEZOUD (Table 2).

The full scalar model is used as the basic model (Model 0) and this article does 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 focusses on the question how method bias can bias the full scalar model with respect to cross-cultural comparisons of socio-cultural integration among non-Western minorities in the Netherlands.

The MGCFA analyses have been conducted with Mplus version 6.11 (Muthén & 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 & 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.[4] 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 & Lewis, 1973) and the comparative fit index (CFI) (Bentler, 1990).

[4] 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.

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 & 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 socio-cultural 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 Roles* and *Familiy 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 socio-demographic variables on the respondent, were regressed on the latent variables and indicators (see Table 2:

Socio-demographic 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 socio-demographic composition (CBS-statline, 2014). A propensity score weighting method is used to investigate how the incomparability of the socio-demographic composition of samples (IoS) between ethnic groups affects cross-cultural comparisons (Bia & Mattei, 2008; DiNardo et al., 1996; Huang et al., 2005; Imbens, 2000; Rosenbaum & Rubin, 1983).

The selection of important socio-demographic 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 socio-demographic background variables have a significant effect on the different categorical indicators (see Table 2: Socio-demographic 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 socio-demographic background variables was conducted. As a third step, only those socio-demographic 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 socio-demographic 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.[5]

[5] 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 socio-cultural 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).

# 3 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 3).[6]

The factor means of *Gender Roles* and *Familiy Ties* of the different ethnic groups are shown in Figures 1 and 2 under M0. Figures 1 and 2 show the change in relative positions of the factor means of *Gender Roles* and respectively *Familiy 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 respective standard errors, see Appendix A. It can be seen that Turkish and Moroccans have, one average, a similar, more traditional attitude towards *Gender Roles* and *Familiy 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 4 and 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 4 and 5). The remaining group comparisons all show significant differences between ethnic groups for both factor means.[7]

### *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 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 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 1 and 2). However, there is no significant difference for *Gender Role* anymore between Moroccans and Antilleans (compare M0 and M1 in Table 4). This means that the observed difference between Moroccans and Antilleans in Model 0 is the result of interviewer effects.

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

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

*Table 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 |

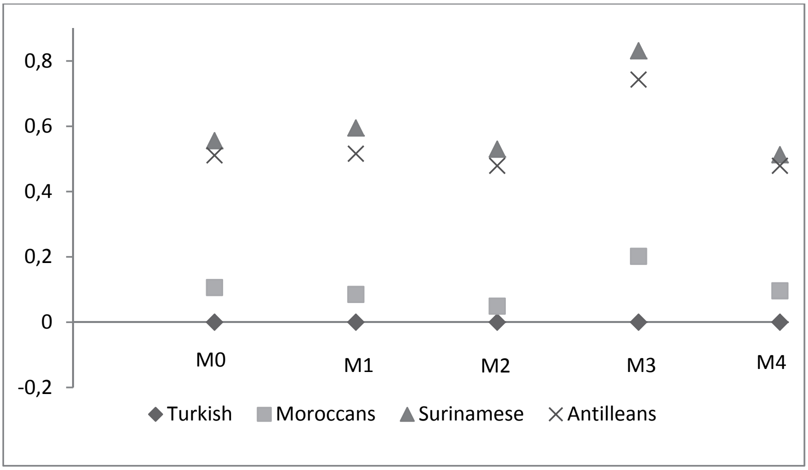

*Figure 1*: Relative positions on Gender Roles of the ethnic groups

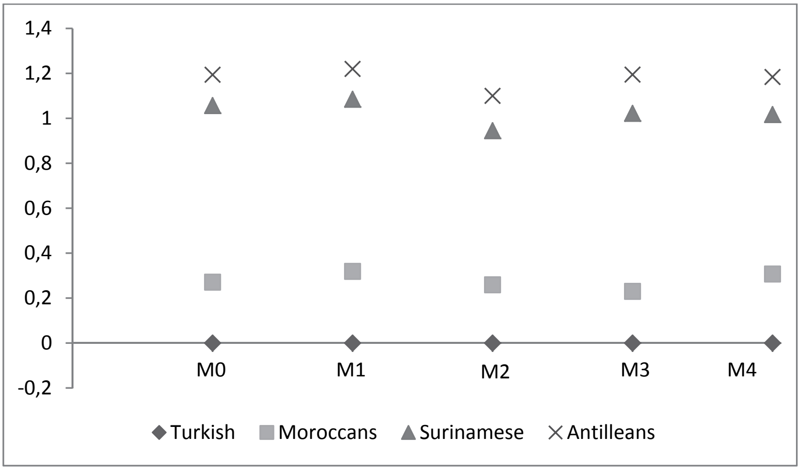

*Figure 2*: Relative positions on Family Ties of the ethnic groups

### *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 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 1 and 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 4 and 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 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 *Familiy Ties*. The complete results can be seen in appendix B. Model 3 (M3) shows an acceptable fit (Table 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 *Familiy Ties*. In all cases more traditional views with respect to *Familiy 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.

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.

*Table 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

*Table 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

*Table 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 *presence of others* during the interview significantly affects the attitudes of Surinamese for both *Gender Roles* and *Familiy 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 *Familiy 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 1 and 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 socio-demographic characteristics were also included as covariates to take into account the nonrandom allocation of these source of method bias. Model 3 (M3) in Tables 4 and 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 socio-demographic sample composition between ethnic groups. A summary of the significant differences between the ethnic groups for eight socio-demographic variables is given in Table 7 (see Table 2 for a description of the socio-demographic variables included in this comparison and Appendix 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.[8] Using the propensity weighting procedure described in section 2.2, only seven of these significant

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

*Table 7*: Summary of the significant differences in socio-demographic 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

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 socio-demographic variables did not have a significant impact on the indicators used to measure *Gender Roles* and *Family Ties* (see also Appendix C).

The comparison of Model 4 (M4) with Model 1 (M1) shows the impact of differences in sample composition for five socio-demographic variables (*Immigration generation, Educational level, Big city dweller, Employed and Children*, see Table 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 3).

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 1 and 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 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 5).

# 4 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 socio-demographic 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 socio-demographic 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 *Familiy 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 *Familiy 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 *Familiy 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 *Familiy 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 *Familiy Ties.* The third research question – how does the language of the interview affect the comparison of attitudes on *Gender Roles* and *Familiy 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 socio-demographic variables such as education, gender, age, etc., was regressed as covariate on the latent variables of *Gender Roles* and *Familiy Ties.* For this a multi group MIMIC (Multiple Indicators MultIple Causes) model was used. The inclusion of the socio-demographic 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 *Familiy 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 *Familiy Ties* and being interviewed in Dutch leads to less traditional views towards *Gender Roles* and *Familiy 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 Turkish and gender match among Moroccans 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 Turkish. Also, Moroccan men have more traditional attitudes towards *Gender Roles* when they are interviewed by a male interviewer compared to the Moroccan men that were interviewed by a female interviewer. Moroccan 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 Turkish respondents and the effect of gender match was only present among Moroccan 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 *Familiy 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 *Familiy 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 *Familiy Ties,* but only among Surinamese and only with respect to *Familiy Ties* among Antilleans.

The modeling of the incomparability of samples was done using a propensity score reweighting procedure of the socio-demographic 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 *Familiy Ties* between non-Western groups can be attributed to differences in socio-demographic 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 *Familiy 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 socio-demographic 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 countries or groups. This can have serious consequences if the survey results contribute towards deciding whether or not a policy is effective in reducing an observed socio-economic or socio-cultural 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 socio-demographic 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 socio-demographic 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 socio-demographic 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.

# References

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 March 2014).

Anderson, B. A., Silver, B. D., & 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, 289-324.

Arends-Tóth, J. & 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.

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

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

Bia, M. & 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.

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

Billiet, J. B. & 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., & Wilhelm, A. (2011). Sampling: The selection of cases for culturally comparative psychological research. In D. Matsumoto & F. J. R. Van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 101-129). New York: Cambridge University Press.

Braun, M. & 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, & T. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 375-393). Wiley Online Library.

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

Chen, C., Lee, S. y., & 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., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research*, 36,4, 462-494.

Cheung, G. W. & 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., & Coulter, P. B. (1982). Race-of-interviewer effects in telephone interviews. *Public Opinion Quarterly*, 46,2, 278-284.

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

Dagevos, J. & Schellingerhout, R. (2003). Sociaal-culturele integratie. Contacten, cultuur en orientatie op de eigen groep. In J.Dagevos, M. Gijsberts, & v. C. Praag (Eds). *Rapportage minderheden*. [In Dutch: Socio-Cultural integration. Contacts, culture and focus on the own ethnic group] Den Haag: SCP, pp. 317-362.

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

Davidov, E., Schmidt, P., & 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., & Resnicow, K. (2010). Interviewer effects in public health surveys. *Health Education Research*, 25, 1, 14-26.

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., & Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrics*, 64,5, 1001-1044.

Fernandez, A. L. & 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., Kappelhof, J., Dagevos, J., & 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.

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

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

Groeneveld, S. & 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. & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

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

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

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

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

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

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., & 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. & Van de Vijver, F. J. (2003). Social desirability in cross-cultural research. In J. Harness, F. J. van de Vijver, & Mohler, P. (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.

Kankaras, M. & 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. (Accepted). The effect of different survey designs on nonresponse in surveys among non-Western minorities in The Netherlands. *Survey Research Methods*, to appear in volume 9, 2, 2014.

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

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

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

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

Lubke, G. H., Dolan, C. V., Kelderman, H., & 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. & 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. & Muthén, B.O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods*, 10, 1, 21-39.

Marin, G., Gamba, R. J., & 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.

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

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

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

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

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

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

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

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

Rosenbaum, P. R. & 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. & 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.

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

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

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 Hague: The Netherlands institute for Social Research/SCP.

Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving survey response: Lessons learned from the European Social Survey*. Chichester, UK: John Wiley & Sons.

Sudman, S. & Bradburn, N. M. (1974). R*esponse effects in surveys: A review and synthesis.* Aldine Publishing Company Chicago, Ill.

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin*, 133, 5, 859-833. Retrieved from American Psychological Association

Tucker, L. R. & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1, 1-10.

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'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*. Vrije Universiteit Amsterdam, Amsterdam.

Van de Vijver, F. J. R. (2003). Bias and equivalence: Cross-cultural perspectives. In J. Harness, F. J. van de Vijver, & Mohler, P. (Eds.), *Cross-cultural survey methods* (pp. 143-155). New York: Wiley.

Van de Vijver, F. J.R. (2011). Capturing bias in structural equation modeling. In E.Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis. Methods and applications* (pp. 3-34). London, England: Routledge.

Van de Vijver, F. & Leung, K. (1997). Methods and data analysis of comparative research.In: Berry, John W.; Poortinga, Ype H.; Pandey, Janak (Eds). *Handbook of cross-cultural psychology*, Vol. 1: Theory and method (2nd ed.). Handbook of cross-cultural psychology (2nd ed.)., (pp. 257-300). Needham Heights, MA, US: Allyn & Bacon, xxv, 406 pp.

Van de Vijver, F.J.R. & 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.

Vandenberg, R. J. & 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.

Veenman, J. (2002). Interviewen in multicultureel Nederland. In: H. Houtkoop en Veenman, J. (Eds), *Interviewen in de multiculturele samenleving. Problemen en oplossingen*. [In Dutch: Interviewing in the multi-cultural Netherlands] Assen: Koninklijke Van Gorcum. pp. 1-19.

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.

Van der Zouwen, J. (2006). De interviewer, hulp of hindernis? In: A.E.Bronner, P. Dekker, E. D. d. Leeuw, L. J. Paas, K. d. Ruyter, A. Smidts, & J. E. Wieringa (Eds.), *Ontwikkelingen in het Marktonderzoek*, Jaarboek 2006 (pp. 63-76). [In Dutch: The Interviewer: help or impediment?] haarlem: spaarenhout.

## Appendix 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) |

*Appendix A continued*

| Parameter estimates (se) | M0 | | M1 | | M2 | | M3 | | M4 | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\lambda^{Ft}_{Vertrffama}$ | 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) |
| $\tau^{Inkjongs}_{2}$ | -0.440 | (0.039) | -0.434 | (0.092) | -0.519 | (0.093) | -0.516 | (0.291) | -0.506 | (0.099) |
| $\tau^{Inkjongs}_{3}$ | -0.101 | (0.038) | -0.120 | (0.094) | -0.192 | (0.092) | -0.148 | (0.287) | -0.194 | (0.107) |
| $\tau^{Inkjongs}_{4}$ | 0.981 | (0.051) | 1.006 | (0.132) | 0.965 | (0.132) | 1.031 | (0.281) | 0.965 | (0.164) |
| $\tau^{Vrwstopw}_{1}$ | -1.473 | (0.059) | -1.457 | (0.128) | -1.564 | (0.127) | -1.451 | (0.277) | -1.606 | (0.123) |
| $\tau^{Vrwstopw}_{2}$ | -0.573 | (0.039) | -0.596 | (0.081) | -0.714 | (0.078) | -0.535 | (0.251) | -0.715 | (0.081) |
| $\tau^{Vrwstopw}_{3}$ | -0.183 | (0.035) | -0.208 | (0.077) | -0.295 | (0.076) | -0.131 | (0.241) | -0.302 | (0.084) |
| $\tau^{Vrwstopw}_{4}$ | 0.940 | (0.047) | 0.925 | (0.126) | 0.883 | (0.130) | 1.059 | (0.242) | 0.869 | (0.145) |
| $\tau^{Thuishuw}_{1}$ | -0.791 | (0.051) | -0.722 | (0.106) | -0.820 | (0.125) | -0.692 | (0.264) | -0.866 | (0.099) |
| $\tau^{Thuishuw}_{2}$ | 0.315 | (0.043) | 0.335 | (0.057) | 0.201 | (0.066) | 0.416 | (0.247) | 0.235 | (0.060) |

*Appendix A continued*

| Parameter estimates (se) | M0 | | M1 | | M2 | | M3 | | M4 | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\tau_3^{Thuishuw}$ | 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) |
| $\chi^2$ | 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. $\lambda^{factor}$ = factorloading of the indicator; $\tau_x$ = 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 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) | | Surinamese (N=779) | | Antilleans (N=852) | |
|---|---|---|---|---|---|---|---|---|
| | GR | FT | GR | FT | GR | FT | GR | FT |
| Intercept | 0.000 (0.000) | 0.000 (0.000) | 0.566 (0.371) | 0.583 (0.432) | 0.404 (0.485) | 1.272 (0.388)* | 0.611 (0.362) | 1.395 (0.348)* |
| Big City Dweller | -0.340 (0.183) | -0.069 (0.098) | -0.198 (0.128) | -0.154 (0.141) | 0.045 (0.141) | -0.113 (0.100) | -0.219 (0.092)* | -0.273 (0.120)* |
| Employed | 0.229 (0.075)* | 0.018 (0.073) | 0.179 (0.066)* | 0.019 (0.177) | 0.182 (0.109) | 0.101 (0.084) | 0.044 (0.068) | 0.059 (0.079) |
| Has Child(ren) | -0.180 (0.171) | -0.388 (0.114)* | 0.110 (0.105) | 0.019 (0.177) | 0.080 (0.093) | -0.071 (0.097) | 0.008 (0.111) | -0.225 (0.096)* |
| Has a partner | 0.050 (0.093) | -0.096 (0.089) | -0.120 (0.092) | -0.260 (0.140) | 0.088 (0.067) | 0.085 (0.073) | 0.064 (0.071) | 0.077 (0.073) |
| Educational level | 0.101 (0.043)* | 0.180 (0.046)* | 0.082 (0.036)* | 0.104 (0.047)* | 0.171 (0.065)* | 0.088 (0.048) | 0.187 (0.047)* | 0.272 (0.052)* |
| Male | -0.232 (0.090)* | 0.176 (0.101) | -0.579 (0.081)* | -0.217 (0.113) | -0.671 (0.145)* | -0.057 (0.074) | -0.604 (0.099)* | -0.080 (0.099) |
| First generation immigrant | 0.032 (0.154) | 0.013 (0.121) | 0.093 (0.125) | -0.192 (0.122) | -0.223 (0.093)* | -0.426 (0.096)* | -0.286 (0.094)* | -0.264 (0.113)* |
| *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) | 0.004 (0.139) | 0.168 (0.130) | 0.004 (0.105) | -0.090 (0.119) |
| 35 – 44 year | 0.162 (0.174) | 0.558 (0.178)* | -0.013 (0.140) | 0.353 (0.288) | -0.153 (0.131) | 0.221 (0.151) | 0.017 (0.116) | -0.007 (0.150) |
| 45 – 54 year | 0.016 (0.189) | 0.554 (0.191)* | 0.004 (0.145) | 0.495 (0.211)* | -0.115 (0.149) | 0.106 (0.137) | 0.075 (0.140) | 0.130 (0.146) |
| 55 – 64 year | 0.069 (0.139) | 0.510 (0.179)* | -0.045 (0.182) | 0.513 (0.286) | -0.066 (0.154) | 0.133 (0.152) | 0.001 (0.127) | -0.219 (0.159) |

*Appendix B continued*

| Parameter estimates (se) | Turkish (N=812) | | | | Moroccans (N=805) | | | | Surinamese (N=779) | | | | Antilleans (N=852) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GR | | FT | | GR | | FT | | GR | | FT | | GR | | FT | |
| 65 + year | -0.075 | (0.221) | 0.344 | (0.232) | -0.115 | (0.183) | 0.331 | (0.262) | -0.147 | (0.183) | -0.061 | (0.161) | -0.135 | (0.182) | -0.060 | (0.229) |
| Others were present | -0.249 | (0.160) | -0.109 | (0.170) | -0.012 | (0.159) | -0.298 | (0.184) | -0.689 | (0.202)* | -0.405 | (0.162)* | -0.062 | (0.119) | -0.259 | (0.100)* |
| Interviewed in native language | -0.142 | (0.100) | -0.364 | (0.131)* | -0.105 | (0.114) | -0.356 | (0.140)* | -0.414 | (0.856) | -0.013 | (0.438) | -0.169 | (0.132) | -0.241 | (0.113)* |
| Gender match interaction | -0.294 | (0.117)* | -0.048 | (0.162) | 0.133 | (0.184) | 0.015 | (0.208) | 0.168 | (0.182) | 0.123 | (0.117) | 0.006 | (0.126) | -0.079 | (0.145) |
| Gender interviewer | -0.022 | (0.157) | -0.043 | (0.156) | -0.339 | (0.159)* | -0.070 | (0.195) | 0.031 | (0.197) | -0.050 | (0.120) | -0.054 | (0.102) | -0.094 | (0.128) |

Note. * = p <0.05

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

*Table C1*: Observed differences on socio-demographic 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 | | | * |

*Table C1 continued*

| | | | | Significant differences between ethnic groups (bonferonni adjusted) | | |
|---|---|---|---|---|---|---|
| Variable | Ethnic group | estimate | se | Turkish | Moroccans | Surinamese |
| 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)

*Table C2*: Observed differences on socio-demographic variables between ethnic groups after propensity score weighting

| | | | | Significant differences between ethnic groups (bonferonni adjusted) | | |
|---|---|---|---|---|---|---|
| Variable | Ethnic group | estimate | se | Turkish | Moroccans | Surinamese |
| Men (proportion) | Turkish | 0.523 | 0.025 | ■ | ■ | ■ |
| | Moroccans | 0.523 | 0.020 | | ■ | ■ |
| | Surinamese | 0.494 | 0.017 | | | ■ |
| | Antilleans | 0.515 | 0.019 | | | |
| Age Group (mean) | Turkish | 2.522 | 0.049 | ■ | ■ | ■ |
| | Moroccans | 2.562 | 0.045 | | ■ | ■ |
| | Surinamese | 3.141 | 0.056 | * | * | ■ |
| | Antilleans | 2.757 | 0.051 | | | * |
| First generation immigrant (proportion) | Turkish | 0.621 | 0.024 | ■ | ■ | ■ |
| | Moroccans | 0.617 | 0.021 | | ■ | ■ |
| | Surinamese | 0.641 | 0.017 | | | ■ |
| | Antilleans | 0.639 | 0.018 | | | |
| Educational level (mean) | Turkish | 2.628 | 0.048 | ■ | ■ | ■ |
| | Moroccans | 2.640 | 0.045 | | ■ | ■ |
| | Surinamese | 2.597 | 0.037 | | | ■ |
| | Antilleans | 2.621 | 0.036 | | | |
| Big City Dweller (proportion) | Turkish | 0.352 | 0.025 | ■ | ■ | ■ |
| | Moroccans | 0.339 | 0.021 | | ■ | ■ |
| | Surinamese | 0.327 | 0.017 | | | ■ |
| | Antilleans | 0.326 | 0.019 | | | |
| Employed (proportion) | Turkish | 0.687 | 0.018 | ■ | ■ | ■ |
| | Moroccans | 0.673 | 0.018 | | ■ | ■ |
| | Surinamese | 0.662 | 0.017 | | | ■ |
| | Antilleans | 0.671 | 0.016 | | | |

*Table C1 continued*

| | | | | Significant differences between ethnic groups (bonferonni adjusted) | | |
|---|---|---|---|---|---|---|
| Variable | Ethnic group | estimate | se | Turkish | Moroccans | Surinamese |
| Has child(ren) (proportion) | Turkish | 0.585 | 0.024 | ■ | ■ | ■ |
| | Moroccans | 0.583 | 0.021 | | ■ | ■ |
| | Surinamese | 0.617 | 0.017 | | | ■ |
| | Antilleans | 0.575 | 0.018 | | | |
| Has partner (proportion) | Turkish | 0.632 | 0.022 | ■ | ■ | ■ |
| | Moroccans | 0.589 | 0.021 | | ■ | ■ |
| | Surinamese | 0.511 | 0.018 | * | * | ■ |
| | Antilleans | 0.499 | 0.018 | * | * | |

Note. * p<0.05/no. of pairwise comparisons. Variables included in the propensity score reweighting: Immigration generation, Educational level, Big city dweller, Employed and Children