Survey Research Methods (2013)
Vol.7, No.1, pp. 17-28
ISSN 1864-3361
http://www.surveymethods.org

# Measurement invariance and quality of composite scores in a face-to-face and a web survey

Melanie A. Revilla
Research and Expertise Centre for Survey Methodology,
Universitat Pompeu Fabra

Measurement equivalence is a pre-requisite to be able to make comparisons across groups. In this paper we are interested in testing measurement equivalence across respondents answering surveys done using different modes of data collection. Indeed, different modes of data collection have specific characteristics that may create measurement non-equivalence across modes. If this is so, data collected in different modes cannot be compared. This would be problematic since, in order to respond to new challenges, like costs and time pressure, more and more often researchers choose to use different modes to collect their data across time, across surveys, and across countries. Studying data about trust and attitudes towards immigration, this paper shows that measurement equivalence holds across a face-to- face and a web survey done in the Netherlands (2008–2009). Moreover, the quality estimates of the Composite Scores are quite high and pretty similar in the two surveys for the four concepts considered.

**Keywords:** measurement equivalence; composite scores; modes of data collection

## 1 Introduction

Measurement equivalence, if it holds, refers to the fact that two individuals with the same true opinion or attitude (or one individual at two occasions) will give the same answer when asked the same question. This may seem obvious but there are in fact a lot of reasons why measurement equivalence might not hold. The answer of the respondents can indeed be affected by other elements than their true opinions: it can be affected by memory limitations, by the amount of effort the respondents invest in answering, by their concentration, by their use of the language, and so on.

Following the terminology of Northrop (1947), a distinction can be made between concepts by postulation (CP) and concepts by intuition (CI). Concepts by postulation are complex concepts that cannot be directly measured but instead are defined by several concepts by intuition. These CPs are represented by latent variables in the models. The concepts by intuition are simple concepts that can be directly measured by items (Saris and Gallhofer 2007). For instance, political trust is a concept by postulation, a broad concept that can be operationalized by identifying and specifying its different components. Thus, political trust can be decomposed into different CIs: trust in the parliament, trust in the legal system, trust in the police, etc. Each of these CIs can be measured by one single question.

Many concepts studied in social sciences are too complex to be measured by single items. Therefore a lot of studies are based on analyses of CPs. Measurement equivalence is usually assessed at this level. But researchers do not always work with latent variables to assess the CPs. They often combine several items (observed answers to the questions) in some kinds of average scores usually called Composite Scores (CS) or Indices (e.g. Anderson, Lepper and Ross 1980; Peterson et al. 1982; Duckworth and Seligman 2006; etc). Composite Scores are combinations of observed scores that are used as shortcuts to measure the CPs of interest. But these CSs are not perfect measures of the CPs. The strength of the relationship between the CP and the CS can be computed: this corresponds to the quality of the CS (Saris and Gallhofer 2007). This quality indicates how much of the observed variance of the CS is explained by the variance of the CP. It provides information about how well the CS measures what one really intends to measure.

Why should we care about measurement equivalence and quality of CS? We should care because it is a pre-requisite to be able to make comparisons between groups. Observed differences can come from true differences or from a lack of measurement equivalence. At the same time, observed similarity does not guarantee that there are no true differences: the true differences can be cancelled out by differences in the measurement leading to similar observed results. So if measurement equivalence is not assessed first, comparative research cannot be trusted.

Measurement equivalence is most often discussed in the frame of crossnational research (e.g. Singh 1995; Steenkamp and Baumgartner 1998). The idea is that countries have different cultures that make people express themselves differently. The typical cliché is that southern countries are much more willing to use extreme words and to be excessive (“fantastic”, “horrible”) while northern countries are famous for their understatements (“not too bad”, “a bit unpleasant”). If people of different countries express themselves in different

Contact information: Melanie A. Revilla, Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra, Barcelona, e-mail: melanie.revilla@hotmail.fr

ways, then two people with the same opinion can choose different answer categories depending on which country they belong to. Besides the culture, problems in translation may also be a threat to measurement equivalence across countries or language groups (Dumka et al. 1996).

However, cross-national research is not the only context where comparisons are made. Comparisons may also be done across groups of respondents with different characteristics (Schulenberg et al. 1988; Tansy and Miller 1997), across surveys, etc. Our interest is in comparisons across modes of data collection. First, focusing on modes of data collection is important because different modes have different characteristics: for a more complete overview, we refer to de Leeuw (2005) or Dillman et al. (2009). Here, we only underline a few elements. One difference is that some modes are self-completed (postal mail, web) whereas in others an interviewer is present (face-to-face, telephone). The presence of the interviewer may lead to higher social desirability bias, i.e. over-reporting of socially desirable attitudes or opinions and under-reporting of the undesirable ones. For example, Kreuter, Presser and Tourangeau (2009) find that web surveys increase the reporting of sensitive information relative to telephone surveys. Since face-to-face surveys also require the mediation of an interviewer, it can be expected that web surveys also increase the reporting of sensitive information relative to face-to-face surveys. Consequently, people with the same true score can pick different answer categories, disturbing measurement equivalence. In particular, the observed means of the variables for socially desirable (respectively undesirable) attitudes are expected to be higher (respectively lower) in presence of an interviewer than in self-completed modes.

Another difference between modes is the kind of stimuli they elicit. Some modes are associated with visual stimuli (postal mail, web) whereas others are associated with oral stimuli (face-to-face, telephone). However, a combination of both visual and oral stimuli is possible (e.g. face-to-face using show cards or web surveys with added voice). Depending on the nature of the stimuli, different ways of answering the questions can be expected. Krosnick (1991) argue that many respondents choose to satisfice, i.e. to minimize their efforts in responding to questions while providing the appearance of compliance. When the answer categories are presented visually, this may lead to primacy effects, which is a bias toward selecting earlier response options instead of considering carefully the entire set of responses. On the contrary, in oral modes, because of memory limitations, respondents are expected to choose more often the last answer categories. This is referred to as "recency effect" (Smyth et al. 1987). Again, this may threaten measurement equivalence across modes.

Secondly, studying equivalence across modes is important because, nowadays, different modes are available to conduct surveys. Each of them has some strengths and weaknesses and it is difficult to say if one is better than the others. It depends on time and costs constraints, on countries' customs for surveys, on the availability of sampling frames, on the coverage of the population for certain modes (e.g. availability of access to the Internet), and on the length of the survey, the topic, its sensitivity, etc. As a result, several modes are regularly used nowadays. Comparing results from surveys using different modes, or results from the same survey at two different points in time after a switch of modes occurred, cannot be done without first assessing if measurement equivalence holds. Besides, some surveys try to solve the problems of low response rates by combining modes within one single survey. In this kind of mixed-mode surveys, it is again crucial to assess measurement equivalence across modes in order to be able to combine the data coming from the different modes.

Finally, there is quite some interest in comparing modes, but usually the focus is on comparing response rates (Hox and De Leeuw 1994; Fricker, Galesic, Tourangeau and Yan 2005) or social desirability bias (Tourangeau and Smith 1996; Kreuter, Presser and Tourangeau 2009). Not much is known about measurement equivalence across modes. King and Miles (1995) look at the measurement equivalence of data collected from paper-and-pencil and computerized formats. Cole, Bedeian and Field (2006), as well as De Beuckelaer and Lievens (2009), test measurement equivalence across paper-and-pencil questionnaires and web surveys. All these analyses find strong support for measurement equivalence, but they are focusing on self-completed modes with only visual stimuli. Does measurement equivalence still hold when an interviewer is present in one mode but not in the other? And when the stimuli are visual in one mode but both visual and oral in another?

The goal of this paper is to investigate whether measurement equivalence holds for different topics in two surveys, one conducted face-to-face in the respondents' house and the other online. The analyses also look at the quality of different composite scores. As far as we know, research on that point is still missing from the literature, so here is a second contribution of our research to the literature. The surveys and topics are presented first, followed by some information about the method, and then the results. Finally, some general conclusions are drawn, together with limits and ideas for future research.

## 2 The surveys and topics

### *2.1 The surveys: European Social Survey (ESS) versus Longitudinal Internet Studies for the Social sciences (LISS) panel*

The comparison is made between two surveys using different modes of data collection, but collecting the data in the same period (end 2008–beginning 2009) in the same country (The Netherlands[1]) and on probability-based samples drawn from a frame of postal addresses.

The first survey is round 4 of the ESS. Many things could be said about this survey[2] but what is most relevant for our

[1] The ESS is conducted in many more countries but we focus on the Dutch data.

[2] More details can be found on the ESS website: http://www.europeansocialsurvey.org/

analyses is that it is a face-to-face survey conducted by an interviewer at the respondent's home and using show-cards. Slightly fewer than 1800 respondents completed the survey in The Netherlands, which corresponds to a response rate of 52%. The second survey is one completed by almost 3200 members of the LISS panel, which is a Dutch web panel.[3] This represents 65.5% of the panel members, and 31.5% of the initial sample. Similar questionnaires were asked to the respondents since the questionnaire proposed to the LISS respondents was adapted from the ESS round 4 questionnaire, keeping constant across modes everything that could be (e.g. same wording of the questions, same scales).

The Netherlands currently have one of the highest Internet penetration rates of Europe, with 88.3% of the population having access to the Internet in 2011.[4] Compared to other countries, its population is in average more web-literate, but it is quite similar to the situation of Nordic countries (e.g. Sweden or Denmark), and within a few years we can expect more countries to present a similar profile. Therefore, it is an interesting country to investigate.

### *2.2 The topics: trust and attitude toward immigration*

Four concepts related to two different topics are used for the comparison. First, the topic of trust has been chosen because many influential scholars, from Hobbes to Weber, defend the idea that trust is essential for social, economic, and political life, at the micro and macro levels. Newton (2007:356) summarizes that: "trusting individuals are said to live longer, happier, and more healthy lives; hightrust societies are said to be wealthier and more democratic; trusting communities are supposed to have better schools and lower crime rates". As a consequence, trust is a central concept for political and social science research. Moreover, trust can be divided into two sub-concepts, social and political trust, because "people may trust those around them and not their political leaders" (Newton 2007:344). Social trust and political trust are complex concepts. These are the first two CPs that we are going to analyse ("soctrust" and "trustin").[5]

The second topic, attitude toward immigration, gained prominence because of the growth of this phenomenon and of the problems related to it: social tensions and conflicts, racism, assimilation of new comers, etc. Most European countries (EU-15) have sizeable immigrant populations today. Consequently, the attitudes of the citizenry towards newcomers have recently been much studied (e.g. Coenders 2001; Mayda 2006). This topic has also been chosen because it is one of the most sensitive topics in the core questionnaire of the ESS round 4. As such, social desirability bias may be expected to be higher in a face-to-face survey than in a web survey (no interviewer). Two concepts related to attitudes toward immigration are present in the ESS and LISS data. The first measures the evaluation of the consequences of immigration: the higher the score of respondents on this variable, the more favourable are their opinions about the impact of immigration. Since the scale goes from negative to positive evaluations, we will call this variable "positive". On the contrary, the second latent variable measures the reluctance of respondents to allow more people to come to the Netherlands. The higher the score on this variable, the less willing people are to accept more immigrants. Therefore, we will call this variable "not allow". These are the third and fourth CPs that we are going to analyse.

Each of these four CPs has several reflective indicators. The CP of social trust has two indicators: how much the respondent thinks people can be trusted and how much he or she thinks that people try to be fair. The three other CPs have three reflective indicators. For political trust, they correspond to the trust in the parliament, in the legal system and in the police. For the evaluation of the consequences of immigration, they correspond to the opinion that immigration is good for the economy, that it enriches culture life, and that it makes the Netherlands a better place to live. Finally, for the reluctance of allowing more people to come and live in the Netherlands, each indicator asks for a different group of immigrants: people from the same race or ethnic group as most Dutch people, people from a different one and people from poorer countries outside Europe.

The names of the variables in the ESS dataset, the wording of the questions and characteristics of the scales can be found in Table 1.

## 3. Method

### *3.1 Testing for measurement equivalence*

This section presents how to test for measurement equivalence (also called invariance) across groups for concepts with reflective indicators. The basic measurement model used is presented in Figure 1.

In this model, $\eta_j$ is the $j^{th}$ latent variable of interest (the CP), the $Y_{ij}$ are the observed variables corresponding to the $i^{th}$ CIs for the $j^{th}$ latent variable of interest, the $\lambda_{ij}$ are the loadings, the $\tau_{ij}$ are the intercepts and the variables $e_{ij}$ represent the random components. It is usually recommended to have at least three indicators for each CP (e.g. see Saris and Gallhofer 2007, Chapter 16). The model can also be expressed with a system of equations:

$$Y_{ij} = \tau_{ij} + \lambda_{ij}\eta_j + e_{ij} \;\; for \;\; all \;\; i, j \qquad (1)$$

Each equation is similar to a regression equation, where $Y_{ij}$ is the dependent variable that one tries to explain, $\eta_j$ is the independent or explanatory variable, $\tau_{ij}$ is the intercept or value of the dependent variable when the independent variable is 0, $\lambda_{ij}$ is the slope, i.e. the increase in $Y_{ij}$ expected for each one unit increase in $\eta_j$, and $e_{ij}$ is the error term. Basic assumptions are made: the error terms are assumed not to

[3] More details can be found on the LISS website: http://www.centerdata.nl/en/MESS

[4] http://www.internetworldstats.com/stats9.htm#eu

[5] It may be argued that in fact the questions cover only a sub-concept of social trust sometimes referred to as "generalised trust" (see for instance Uslaner 2002) and only a sub-concept of political trust that could be called "trust in institutions", but for simplification purposes, we will call them "social" and "political" trust.

*Table 1:* Experiments about trust and immigration

| Concept | Variable | Meaning | Method |
|---|---|---|---|
| soctrust | ppltrst | - Generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people? | 11 points (from negative to positive) |
| | pplfair | - Do you think that most people would try to take advantage of you if they got the chance, or would they try to be fair? | |
| trustin | | How much do you personally trust each of the institutions: | 11 points (no trust to complete trust) |
| | trstprl | - Dutch parliament | |
| | trstlgl | - The legal system | |
| | trstplc | - The police | |
| positive | imbgec | - It is generally bad for the Dutch economy that people come to live here from other countries | 11 points (from negative to positive) |
| | imueclt | - Dutch cultural life is generally undermined by people coming to live here from other countries | |
| | imwbcn | - The Netherlands are made a worse place to live by people coming to live here from other countries | |
| not allow | imsmet | - The Netherlands should allow more people of the same race or ethnic group as most Dutch people to come and live here. | 4 points (allow many to allow none) |
| | imdfctn | - The Netherlands should allow more people of a different race or ethnic group from most Dutch people to come and live here. | |
| | impcntr | - The Netherlands should allow more people from the poorer countries outside Europe to come and live here. | |

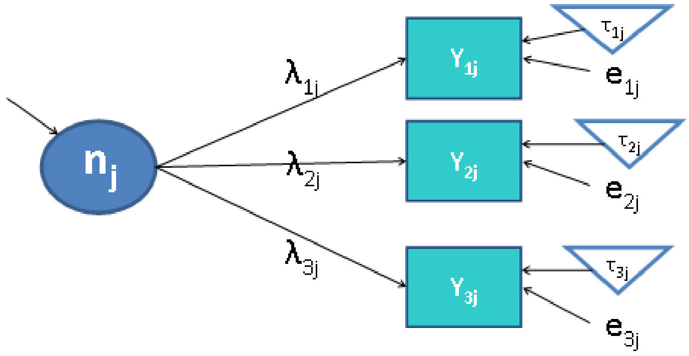

*Figure 1.* The basic measurement model (Note: $\eta_j$ is the $j^{th}$ latent variable; $Y_{ij}$ are the $i'h$ observed variables for the latent trait $j$)

be correlated with the independent variables, nor with each other. The different latent variables ($\eta_j$) are assumed to be correlated with each other.

In order to fix the scale of the latent variables, for each CP, one of the loadings, e.g. the one of the first observed variable ($\lambda_{lj}$), is fixed to 1 and one of the intercepts, e.g. $\tau_{lj}$, is fixed to 0.

The same model is specified in the different groups that one wants to compare: in our case, the face-to-face and the web surveys. Using a multiple-group confirmatory factor analysis approach, it is possible to test for different levels of equivalence, by putting more or less constraints of equality on the parameters across groups.

We sequentially test for the three more common levels of invariance (Meredith 1993):

- configural invariance: the same measurement model holds in all groups (i.e. in the different modes)
- metric invariance: configural invariance holds and the slopes $\lambda_{ij}$ are equal in all groups
- scalar invariance: metric invariance holds and the intercepts $\tau_{ij}$ are the same in all groups

If metric invariance holds, the comparison across groups of the unstandardized relationships between variables is allowed. If scalar invariance holds, the comparison across groups of the means of the CPs is allowed.

The analyses are done in LISREL (Jöreskog and Sörbom 1991) using the Maximum Likelihood estimator for multi-group analyses[6] and analysing the covariance matrices. Pearson's correlations matrices[7], standard deviations and means

[6] Lisrel input available online: http://bit.ly/e2wwpT

[7] Bollen and Barb (1981) show that "when as few as five categories are used to approximate the continuous variables, the correlation coefficients and their standard deviations for the collapsed and continuous variables are very close" (p. 232)

are specified as the input data. One tricky but crucial step is to assess the fit of the model. There are two main ways of looking at a model's fit.

First, one can consider the global fit of the model, using for instance the chi-square test. However, this test has important limits: it is dependent on the sample size, on the size of the parameters, and on the power, it is sensitive to deviations from normality, etc. That is why a huge range of fit indices have been developed lately (RMSEA, CFI, etc), but they have limits too. Nevertheless, some authors (e.g. Cheung and Rensvold 2002; Byrne and Stewart 2006; Chen 2007) argue that it is still possible to use these fit indices to test for measurement equivalence, but focusing on the *changes* in these measures when adding the constraints at the different steps. They consider that a change larger than .01 is an indication of non equivalence. We will therefore look at the changes in RMSEA and CFI for our different models.

However, Saris, Satorra and Van der Veld (2009) argue that there is no proper way to test a model as a whole and that it is necessary to make the test at the parameter level. Following them, the second option is to consider the local fit of a model. We will mainly focus on this way of testing. Indeed, this approach is more adequate for our purpose: that is to test the equality of given parameters of the model (the loadings, the intercepts) and not only of the model as a whole. Besides, the procedure developed by Saris, Satorra and Van der Veld (2009) also takes into account the power. Therefore, by using JRule software (Van der Veld, Saris, Satorra 2009) based on their procedure, we are able to test for specific equalities in our model and take not only type I but also type II errors into account. This software considers the modification indices, the power and when necessary the expected parameter changes in order to determine if, and where, there are misspecifications in the model (see Appendix Table 1). It suggests how the model can be corrected to improve its fit.

We should notice however that what is considered as a misspecification depends on what the researcher wants to detect: if he/she wants to detect a deviation of $x$, JRule tells him/her where there are deviations higher than $x$. This is what is referred to as misspecifications. We used the following values to define a misspecification: 0.10 for loadings, 0.10 for causal effects and correlations, 0.03 times the scale range for the intercepts and mean structure.[8]

Measurement invariance informs us about the possibility of comparing unstandardized relationships and means across groups. Even if scalar invariance holds, however, the standardized estimates can be different across surveys if the variances vary. We therefore consider in the next sections the quality of the CSs, an indicator based on standardized parameters, and the correlations between the two trust concepts on a one hand and between the two immigration concepts on the other hand, which gives us some clue about external validity.

## 3.2 Computing the quality of the composite scores

Two different kinds of CS are generated using Stata version 10 (StataCorp 2007). First, we generate what we call the "basic" CSs, which are unweighted averages of the different questions that are part of them ($w_i$ = 1 / number of indicators). We are interested in the unweighted approach because it is the most widely used by researchers. However, more elaborated weights can be used as well. Therefore, we also generate CSs using regression weights: these weights minimize the sum of squared differences in scores between the CP and the CS (Saris and Gallhofer 2007:283). They should be computed on the pooled data (putting together the different groups), otherwise, differences can be found that come from the difference in weights.

We use LISREL to generate these regression weights. For three out of the four concepts that we are studying, we estimate on the pooled data a simple factor model with one latent variable and three observed indicators in order to get the regression weights (we just need to ask for the "factor scores" in the input and LISREL provide the regression weights automatically). For social trust, we only have two indicators. The model, therefore, is not identified. In order to get some weights, we estimate a factor model including social trust together with political trust. Since they are correlated, the model is identified and we can get regression weights. The problem is that the weights for one concept can be affected by the indicators of the second concept, because the concepts correlate, so the weights obtained may not be optimal for each concept separately. However, it may still be a better procedure than taking equal weights for the different indicators.[9]

The quality of the CSs can be defined (Lawley and Maxwell 1971; Saris and Gallhofer 2007) in the same way as the quality of single items: it is the strength of the relationship between the CS and the latent variable of interest (CP).

The model in Figure 1 can be extended to include the CS, as shown in Figure 2. The intercepts and error terms have not been explicitly specified, but the small arrows represent them.

The quality, or strength of the relationship between the latent variable ($\eta_j$) and the CS, can be computed as the correlation squared between the latent variable of interest and the CS. For the exact formula and details on the procedure, we refer to Saris and Gallhofer (2007:284).

Discussing the significance of the differences in quality of the CSs across the two surveys is a bit tricky since a formal test would require computing the standard errors of the quality estimate, which is quite complex. Instead, we focus on the relevance of the difference. We consider that a difference in quality of the CSs across surveys is relevant if it changes significantly (0.10 or more, criterion used in JRule)

[8] The default values proposed by JRule are based on what is often used in practice. For instance, in the literature, it is often seen that if a loading is lower than 0.40, it is ignored by the researchers. However, we thought that the default values were too soft, so we changed them to have a stricter test.

[9] A linear transformation is applied to all weights obtained in LISREL in order to get weights whose sum equals to one. These weights are used to compute the CSs.

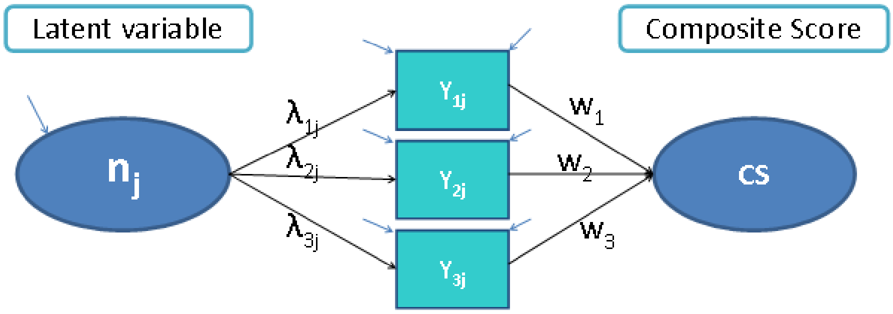

*Figure 2*. Extension of the model to the Composite Score (Note: $\eta_j$ is the $j^{th}$ latent variable; $Y_{ij}$ are the observed variables; $\lambda_{ij}$ the loadings; $w_i$ is the $i_{th}$ weight; the arrows represent the intercepts and error terms.)

the observed correlations we get when the true correlation is the same.

Both the unweighted CSs and the CSs based on regression weights are considered. For three out of four concepts, the values used for the loadings ($\lambda_{ij}$) are the ones obtained in LISREL by running a simple factor model for one concept with three indicators separately for each survey. For social trust, as the regression weights are taken from the combined analysis of this concept together with political trust, the loadings are also taken from such a combined analysis, but estimated separately for the ESS and the LISS.

### *3.3 External validity*

Different types of validity can be distinguished. We focus here on what is called "criterion-related validity" or "external validity". In Alwin's (2007:23) terms: "Criterion-related validity refers to the predictive utility of a measure or set of measures – do they predict or correlate with other theoretically relevant factors or criteria? For example, the criterion-related validity of SAT scores are typically assessed in terms of their ability to predict college grades (Crouse and Trusheim 1988)". Or a few pages later: "Criterion validity is simply defined as the correlation of the measure Y with some other variable, presumably a criterion linked to the purpose of measurement" (Alwin 2007:47).

In our case, the criterion validity is quantified by looking at the correlation between the two trust CPs and at the correlation between the two immigration CPs. The more similar this correlation is to the expected value, the better the external validity. We have to acknowledge that our test of external validity is quite weak, since we do not have a gold standard to correlate to our concepts of interest here but only another variable measured using the same data, and that could therefore suffer from the same drawbacks. So we should be careful about the conclusions we can reach from this simple test.

Another limit is that we cannot directly test if the correlations are significantly different in one mode than in the other. But LISREL allows us to test if the covariances are significantly different, by adding a constraint of equality on these specific parameters. Then we look in JRule whether the program indicates a misspecification for these parameters when they are constrained to be the same. If not, we conclude that the covariances are not significantly different across modes. The correlations can still differ if the variances do, but we expect these differences to be relatively small.

### *3.4 Application*

The model presented in Figure 2 is applied to the two topics of interest, trust and attitude towards immigration, in the way described in Figure 3.

In one given survey (ESS or LISS) and for one given topic (trust or immigration), the model is composed of two latent variables that each has several reflexive indicators. These same items are also used to create the CSs (both for the unweighted model and using regression weights). The CSs are called: "$CS_{soctrust}$", "$CS_{trustin}$", "$CS_{positive}$" and "$CS_{notallow}$" since they respectively intend to measure the CPs of social trust, trust in politics, evaluation of the consequences of immigration and reluctance towards allowing more immigrants. The external validity is tested by looking at the correlation between the two latent variables for one given topic.

For trust, although empirical research does not always find a correlation significantly different from zero between these two CPs (Newton 2007), and although it is necessary to make a distinction between them, theoretically it makes sense to argue that people that tend to trust other individuals also tend to have higher trust in politics such that some positive correlation should be found between them.

For immigration, it is expected that respondents thinking that immigration has negative consequences for the country will also be reluctant to allow more immigrants to come and live in the Netherlands, whereas respondents thinking immigration has positive consequences will be favourable towards allowing more immigrants. The two concepts should therefore be negatively correlated.

## 4 Results

### *4.1 Measurement equivalence*

First, testing for configural invariance, no misspecifications were detected by JRule, so for both topics, the same model holds in the face-to-face and the web surveys. We can notice that although several or all items are measured with the same method, the testing of the model does not suggest

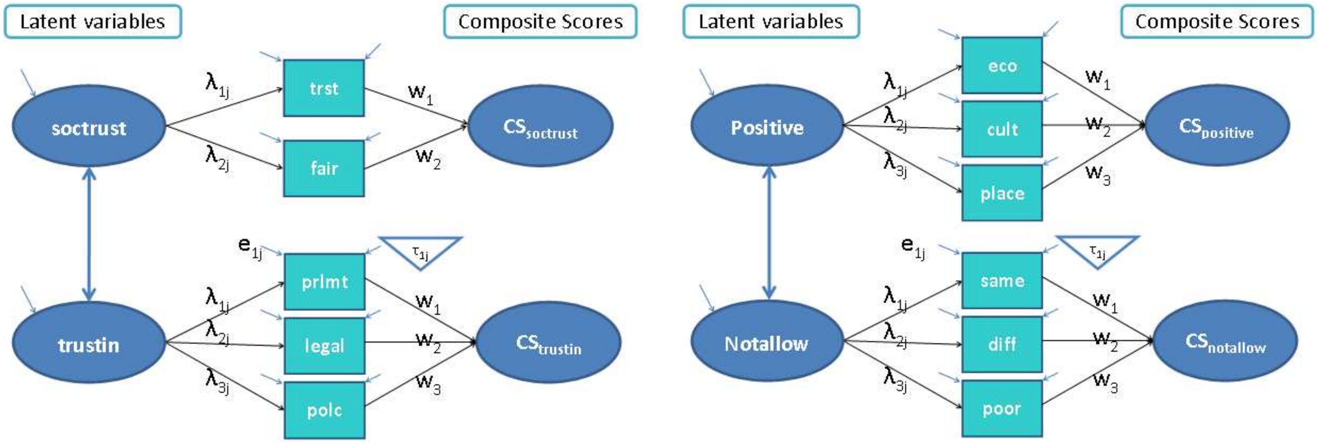

*Figure 3*. Model for the trust example and the immigration example

that we need to introduce a method factor, since no misspecifications are detected, in particular, no correlation between the error terms are suggested.

Since configural invariance holds, we went on with testing the second level of invariance, which is metric invariance (equal slopes). For trust as well as for attitudes towards immigration, JRule does not indicate any misspecification for the parameters constrained to be equal across surveys, i.e. the loadings. In addition, the power is 0.99 in most cases, which means that there is a 99% chance that the test will detect the misspecification if the true difference for one parameter is bigger than the minimal difference we want to detect.

Also, the CFI does not change when adding the constraint of equal loadings in any of the two models (trust and immigration). The RMSEA goes from .059 to .050 in the case of trust and from .068 to .062 in the case of immigration. Both changes are therefore lower than .01.

Overall, according to both testing procedures, metric equivalence cannot be rejected: unstandardized relationships between variables can be compared across the ESS and the LISS surveys for the topics of trust and attitudes toward immigrants.

Finally, scalar invariance is tested by adding equality constraints on the intercepts. Again, JRule does not indicate any misspecification for the parameters of interest (loadings and intercepts now) although there is high power (0.99). Again, the CFI does not change at all. The RMSEA does not change either in the case of immigration. For trust it goes from .050 to .063. This is higher than the threshold suggested for misspecification. But it is however still quite close and the other testing procedure as well as the change in CFI (or absence of change) do not suggest that we should reject scalar invariance, so we consider that scalar invariance holds in that case too.

Since scalar invariance holds, it is possible to compare the means of the CPs. So far, we allowed them to be free in each survey. This led to very similar but not equal means of the latent variables in both surveys, as can be seen in Table 2. In order to see if the differences are statistically significant, we add the constraint in LISREL that they should be invariant across surveys. Using JRule again, even if the power is very high, we cannot reject this hypothesis, meaning that the means of the CPs studied are not significantly different in the two surveys. Their values with additional constraint on the means of the CPs (and model for scalar invariance) are given in Table 2 too.

We start with attitudes toward immigration. Table 2 shows that whatever the mode of data collection used, the CP's means is around 2.3 for "not allow" (measured on a 4-point scale) and 5.3 for "positive" (11 point scale): on average the Dutch population thinks that some or a few more people should be allowed to come and live in the country and that immigration is positive for the country across different domains. However, on average the Dutch population is almost neutral (close to the middle of the scale).

For trust, the means of the latent variables are around 5.9 for both "soctrust" and "trustin" (both measured on an 11-point scale): so the means social and political trust in the Netherlands are in the positive half of the scale, but again close to the center.

The means of the CSs can differ from the means of the CPs, but since scalar invariance holds and the means of the CPs are equal in the LISS and the ESS, the means of the CSs should be similar across the two surveys. So even if the surveys use different modes of data collection, the means of both the CPs and the CSs can be compared across the ESS and LISS.[10]

[10] If each observed variable $Y_{ij}$ is a linear function of the CP, the mean of the unweighted CS is: $E(CS) = (1/3)E(Y_{11} + Y_{21} + Y_{31}) = (1/3)(\lambda_{11} + \lambda_{21} + \lambda_{31})E(\eta_1) + E(\tau_{11} + \tau_{21} + \tau_{31}) + E(e_{11} + e_{21} + e_{31})$. If we assume that the mean of the error terms is 0 and if scalar invariance holds, then a difference in the means of the CS can only comes from a difference in the means of the CPs. If the means of the CPs are equal across groups, then the CSs should also be equal across groups. Still the means of the CSs may vary from the means of the CPs, for instance if the sum of the loadings is different from 3.

*Table 2:* Means of the CPs in both surveys

| | Immigration | | | | Trust | | | |
|---|---|---|---|---|---|---|---|---|
| | Not allow | | Positive | | Soctrust | | Trustin | |
| | ESS4 | LISS | ESS4 | LISS | ESS4 | LISS | ESS4 | LISS |
| Mean Latent variable unconstrained | 2.23 | 2.35 | 5.43 | 5.25 | 5.93 | 5.80 | 5.93 | 5.87 |
| Mean Latent variable adding constraint equality | 2.31 | | 5.31 | | 5.87 | | 5.90 | |

*Table 3:* Regression weights for the pooled data of both surveys

| | Trait 1 | Trait 2 | Trait 3 |
|---|---|---|---|
| "not allow" | .21 | .70 | .09 |
| "positive" | .31 | .40 | .29 |
| "soctrust" | .54 | .46 | na |
| "trustin" | .21 | .61 | .18 |

## *4.2 Quality of Composite Score*

The regression weights obtained on the pooled data of both surveys for the different CSs are presented in Table 3. Table 4 gives the quality of both the unweighted and regression-weights based CSs.

First, if we compare in Table 4 the quality estimates for the basic and the more elaborated CSs, it seems not to matter much. There is only a difference for "not allow", but it is quite small (0.04) and it is the same in both surveys.

For "not allow" the quality is quite high and very similar in both surveys (around 0.90 for the basic CSs and 0.95 for the ones based on regression weights). For the other concepts, the quality is not so high but it is still higher than 0.75. Besides, the differences are larger but they stay small. In order to determine the relevance of these differences, we use the estimates of quality found in the different surveys to examine what differences in observed correlations appear, given a true correlation, due to a variation in quality across surveys. We focus on the topic of trust and on the basic CS since it is there that the greatest differences between the face-to-face and the web surveys are found (0.09 for "trustin" and 0.03 for "soctrust"). The observed correlation between the CS for social and political trust can be expressed as the product of the true latent correlation times the quality coefficient for the CS of social trust times the quality coefficient for the CS of political trust, that is:

$$r(CS_{soctrust}, CS_{trustin}) = \rho \times q_{cs_{soctrust}} \times q_{cs_{trustin}} \qquad (2)$$

where $\rho$ is the true correlation between the CPs of social trust and political trust. So the difference between observed correlations across surveys is:

$$r(CS_{soctrust}, CS_{trustin})^{LISS} - r(CS_{soctrust}, CS_{trustin})^{ESS} = \rho(q^{LISS}_{cs_{soctrust}} q^{LISS}_{cs_{trustin}} - q^{ESS}_{cs_{soctrust}} q^{ESS}_{cs_{trustin}}) \qquad (3)$$

The difference is a linear function of the true correlation: the higher the true correlation, the higher the difference in observed correlations. So in order to see the maximum difference we take a correlation of one. Then, the highest difference for trust is still lower than the value we set as criterion for misspecification. Indeed, we have:

$$r(CS_{soctrust}, CS_{trustin})^{LISS} - r(CS_{soctrust}, CS_{trustin})^{ESS} =$$

$$1 \times (.88 \times .95 - .86 \times .89) = 0.0706$$

$$\text{So } r(CS_{soctrust}, CS_{trustin})^{LISS} - r(CS_{soctrust}, CS_{trustin})^{ESS} =$$

$$= .10$$

In the next section, we will see that the true correlation between the concepts by postulation of social and political trust is around 0.50 (cf. Table 5). If this is so, it means, knowing the quality of the two CSs, that we expect an artificial difference of 0.0706/2 = 0.0353 in the observed correlations of the face-to-face and the web surveys. This is small enough to not worry about. For the other cases (other topic and/or CSs based on regression weights), the differences are even lower.

All in all, the similarities dominate: the quality of the different CSs is close enough in the face-to-face and web surveys for the different concepts to not disturb the cross-survey analyses of standardized relationships. It seems also that basic CSs can perform almost as well as more elaborated CSs.

## *4.3 External validity*

The last result we want to stress concerns the external validity of the four concepts analysed. As argued previously, for the two concepts about immigration, "positive" and "not allow", we assume a negative and quite strong correlation. On the contrary, for the topic of trust, a positive correlation is expected between social and political trust. This correlation should not be too close to 1 (otherwise it would mean that social and political trust are the same concept, which is not what the literature shows). It may even be relatively low, but still it should be significant and positive.

We start with the models for scalar invariance and we run them again but constraining the parameter of the covariance between the two CPs to be the same in the face-to-face and the web survey. This does not lead to misspecification according to JRule, suggesting that the covariances are not signigicantly different across the two modes. However, our interest here is to consider the standardized relationships and

*Table 4:* Quality of the Composite Scores

| | Immigration | | | | Trust | | | |
|---|---|---|---|---|---|---|---|---|
| | Not allow | | Positive | | Soctrust | | Trustin | |
| | ESS4 | LISS | ESS4 | LISS | ESS4 | LISS | ESS4 | LISS |
| $q^2_{basic\ \ CS}$ | .90 | .90 | .77 | .82 | .75 | .78 | .79 | .88 |
| $q^2_{reg-weights\ \ CS}$ | .94 | .95 | .77 | .83 | .75 | .78 | .83 | .91 |
| Absolute diff. between basic and reg-weights CS | .04 | .05 | .00 | .01 | .00 | .00 | .04 | .03 |
| Absolute difference ESS-LISS for basic | .00 | | .05 | | .03 | | .09 | |
| Absolute difference ESS-LISS for reg-weights | .01 | | .06 | | .03 | | .08 | |

*Table 5:* Testing for external validity

| | Immigration | | Trust | |
|---|---|---|---|---|
| | ESS4 | LISS | ESS4 | LISS |
| Corr($CP_1$, $CP_2$) | -.64 | -.64 | .47 | .52 |
| Corr($CS_1$, $CS_2$) basic | -.54 | -.57 | .40 | .42 |
| Corr($CS_1$, $CS_2$) reg-weights | -.54 | -.57 | .38 | .41 |
| External validity | ok | ok | ok | ok |

not the unstandardized ones. Therefore, Table 5 presents for the two topics the correlations between CPs, as well as the correlations between the two composite scores, unweighted and using regression-weights.

One can notice that differences in the correlations for the CPs are found: for the topic of trust, although the unstandardized parameters are equal, the standardized ones vary. Nevertheless, the difference found across surveys is lower than the criterion used to specify a misspecification (0.05<0.10), so we conclude that even if one survey is using face-to-face interviews whereas the other use the web, this does not impact significantly the correlation between the two CPs considered in each of the two topics.

Moreover, the correlations are in line with what we expected for both topics. Indeed, the correlation between "positive" and "not allow" is -0.64 in the ESS and in the LISS: so it is a quite strong negative correlation. On the contrary, the correlation between social and political trust is positive: 0.47 in the ESS and 0.52 in the LISS. This is quite high compared to some past results (Newton 2007) but in line with others (Saris and Gallhofer 2007), and this is still much lower than 1 and seems quite probable. Overall, the analyses suggest that the external validity for the CPs is similar in the ESS and in the LISS surveys, when face-to-face is used and when web is used. We should however be careful about this result since our test of external validity presents many limits. Further research using a gold standard to correlate with the CPs would be necessary to test better external validity of our concepts.

Now, looking at the correlation between the CSs, more differences are observed. These correlations are also always lower in absolute values than the ones between CPs. However, the correlations are all in the expected direction and relatively large (between .38 and .57 in absolute values), suggesting that external validity also applies to the CSs, weighted or not, but here again more research would be needed.

## 5 Conclusion

Measurement equivalence needs to be assessed in order to be able to make meaningful comparisons across groups. In this paper we were interested in groups of respondents that are completing surveys using different modes of data collection. Modes have a set of properties (interviewer or self-completion, visual or oral stimuli) that may influence the way people express themselves when answering a survey. Thus, there is a risk that the mode of data collection could threaten the measurement equivalence of the questions.

Comparing a face-to-face survey (ESS) and a web survey (LISS) for four different concepts related to the two topics of trust and attitude towards immigration, we found that configural, metric and scalar invariances all hold across the two surveys and for all four concepts. Since metric invariance holds, one can compare the unstandardized relationships of the concepts with each other across modes. That scalar invariance holds too suggests that one can also compare the means of these four concepts across different modes of data collection.

But scalar invariance only tells us about unstandardized relationships. Standardized relationships may still vary. For standardized estimates to be comparable we need to have in addition to metric invariance also that the variances of the latent variables are the same. The quality estimates of the CSs, since they are computed using standardized estimates, may therefore vary across surveys even if metric and scalar invariance has been assessed if there are differences in the variances of the factors. However, our results show that the quality estimates are comparable across surveys. Therefore, we can compare standardized measures across the ESS and the LISS for the concepts tested. We also find that using a basic CS or one based on regression weights does not really make a difference.

We looked at one particular standardized measure to illustrate this concept, the correlation between the two CPs of interest. The correlations are equivalent across surveys and the external validity seems to hold too since the correlations found between the two CPs within each topic go in the expected direction and are relatively large.

The analysis however focuses on only four concepts about two topics, considers only two modes, and is based on data from only one country, the Netherlands. Therefore, much more evidence would be needed before being able to generalise our conclusions. Still, overall, the results are quite encouraging, since they show that even using different modes of data collection, as long as the exact same wording and scales are used and the samples are drawn randomly from the population, equivalent measurements can be obtained and CSs of similar and quite high quality can be constructed using the data. The use of show cards in the face-to-face survey is probably also an important element explaining the similarity across the two surveys studied, as well as the fact that the LISS panel provides access to a computer and Internet to the respondents that do not have it. Indeed, even if the LISS study has a quite lower response rate, there is a high similarity with respect to gender, age and education distribution in the two surveys (see Appendix Table 2), which we think is clearly a result of the specific procedure used in the LISS panel to recruit and keep active their panel members. It might still be that our samples differ with respect to other variables that we did not consider, limiting the scope of our results. However, our results are in line with previous research about measurement equivalence across different modes (King and Miles 1995; Cole, Bedeian and Field 2006; De Beuckelaer and Lievens 2009; Davidov and Depner 2011), which suggest they are more general that this study only.

## Acknowledgements

I am extremely grateful to Willem Saris and some anonymous reviewers for their very helpful comments on previous drafts of this paper. Finally, I thank Peter Lynn for his comments as the co-director of my thesis.

## References

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. Hoboken NJ: Wiley.

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of Social Theories: The Role of Explanation in the Persistence of Discredited Information. *Journal of Personality and Social Psychology*, *39*(1-6), 1037-1049.

Bollen, K. A., & Barb, K. H. (1981). Pearson's r and Coarsely Categorized Measures. *American Sociological Review*, *46*(2), 232-239.

Byrne, B. M., & Stewart, S. M. (2006). The MACS Approach to Testing for Multigroup Invariance of a Second-Order Structure: A Walk Through the Process. *Structural Equation Modeling*, *13*(2), 287-321.

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indices to Lack of Measurement Invariance. *Structural Equation Modeling*, *14*, 464-504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255.

Coenders, M. (2001). *Nationalistic Attitudes and Ethnic Exclusionism in a Comparative Perspective: An Empirical Study of Attitudes Toward the Country and Ethnic Immigrants in 22 countries*. Radboud University Nijmegen.

Cole, M. S., Bedeian, A. G., & Field, H. S. (2006). The Measurement Equivalence of web-Based and Paper-and-Pencil Measures of Transformational Leadership: A Multinational Test. *Organizational Research Methods*, *9*, 339-368.

Crouse, J., & Trusheim, D. (1988). *The case against the SAT*. Chicago, IL: University of Chicago Press.

Davidov, E., & Depner, F. (2011). Testing for measurement equivalence of human values across online and paper-and-pencil surveys. *Quality & Quantity*, *45*(2), 375-390.

De Beuckelaer, A., & Lievens, F. (2009). Measurement Equivalence of Paper-and-Pencil and Internet Organisational Surveys: A Large Scale Examination in 16 Countries. *Applied Psychology: an international review*, *58*(2), 336-361.

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

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., et al. (2009). Response Rate and Measurement Differences in Mixed Mode Surveys Using Mail, Telephone, interactive Voice Response and the Internet. *Social Science Research*, *38*(1), 1-18.

Duckworth, A. L., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology*, *98*(1), 198-208.

Dumka, L. E., Stoerzinger, H. D., Jackson, K. M., & Roosa, M. W. (1996). Examination of the cross-cultural and cross-language equivalence of the parenting self-agency measure. *Family Relations*, *45*, 216-222.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An Experimental Comparison of web and Telephone Surveys. *Public Opinion Quarterly*, *69*, 370-392.

Hox, J. J., & De Leeuw, E. D. (1994). A Comparison of Nonresponse in Mail, Telephone, and Face-to-Face Surveys: Applying Multilevel Models to Metaanalysis. *Quality and Quantity*, *28*, 329-344.

Jöreskog, K. G., & Sörbom, D. (1991). *LISREL VII: A guide to the program and applications*. Chicago, IL: SPSS.

King, W. C., & Miles, E. W. (1995). A quasi-experimental assessment of the effect of computerizing noncognitive paper-and-pencil measurements: A test of measurement equivalence. *Journal of Applied Psychology*, *80*(6), 643-651.

Kreuter, F., Presser, S., & Tourangeau, R. (2009). Social Desirability Bias in CATI, IVR and web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly*, *72*(5), 847-865.

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, *5*, 213-236.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*. London: Butterworth.

Mayda, A. M. (2006). Who Is Against Immigration? A Cross-Country Investigation of Individual Attitudes toward Immigrants. *The review of Economics and Statistics*, *88*(3), 510-530.

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

Newton, K. (2007). Social and Political Trust. In R. J. Dalton & H.-D. Klingemann (Eds.), *The Oxford Handbook of Political Behavior* (p. 342-359). Oxford: Oxford University Press.

Northrop, F. S. C. (1947). *The Logic of the Sciences and the Humanities*. New York: World Publishing Company.

Peterson, C., Semmel, A., von Baeyer, C., Abramson, L. Y., Metalsky, G. I., & Seligman, M. E. P. (1982). The attributional style questionnaire. *Cognitive Therapy and Research*, *6*, 287-300.

Saris, W. E., & Gallhofer, I. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. New York: Wiley.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural equation modeling: A multidisciplinary Journal*, *16*(4), 561-582.

Schulenberg, J. E., Shimizu, K., Vondracek, F. W., & Hostetler, M. (1988). Factorial invariance of career indecision dimensions across junior high and high school males and females. *Journal of Vocational Behavior*, *33*, 63-81.

Singh, J. (1995). Measurement issues in cross-national research. *Journal of International Business Studies*, *26*, 597-619.

Smyth, M. M., Morris, P. E., Levy, P., & Ellis, A. W. (1987). *Cognition in Action*. London: Erlbaum.

StataCorp. (2007). *Stata Statistical Software: Release 10*. College Station, TX: StataCorp LP.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in crossnational consumer research. *Journal of Consumer Research*, *25*, 78-90.

Tansy, M., & Miller, J. A. (1997). The invariance of the self-concept construct across White and Hispanic student populations. *Journal of Psychoeducational Assessment*, *15*, 4-14.

Tourangeau, R., & Smith, T. W. (1996). Asking Sensitive Questions: the Impact of Data Collection Mode, Question Format, and Question Context. *Public Opinion Quarterly*, *60*(2), 275-304.

Uslaner, E. M. (2002). *The moral foundations of trust*. New York: Cambridge University Press.

Van der Veld, W. M., Saris, W. E., & Satorra, A. (2009). *Judgement Rule Aid software.*

# Appendix

*Table 1:* Testing procedure with JRule

| Testing in Jrule | Low Power (<.8) | High Power (≥.8) |
|---|---|---|
| Insignificant MI | Inconclusive | No misspecification |
| Significant MI | Misspecification | Inspect EPC |

*Table 2:* Composition of the samples (percent)

| | | ESS4 | LISS study | LISS panel |
|---|---|---|---|---|
| Gender | Men | 46.0 | 44.6 | 49.4 |
| | Women 54.0 55.4 50.6 | | | |
| Age | 16-19 | 4.4 | 2.7 | 7.3 |
| | 20-39 | 28.8 | 27.5 | 32.7 |
| | 40-64 | 45.5 | 52.3 | 49.4 |
| | 65-79 | 17.0 | 15.5 | 10.0 |
| | >80 | 4.3 | 1.9 | 1.0 |
| Education | Low | 37.7 | 35.7 | 33.0 |
| | Middle | 35.6 | 33.2 | 36.9 |
| | High | 26.8 | 31.1 | 30.1 |