Measurement equivalence in multilingual comparative survey research

Diana Zavala-Rojas

TESI DOCTORAL UPF / 2016

Directors de la tesi

Dr. Willem Saris
Dra. Verónica Benet-Martínez

DEPARTAMENT DE CIÈNCIES POLÍTIQUES I SOCIALS

Universitat Pompeu Fabra Barcelona

A la memoria de mi abuelita María Luisa
To the memory of my dear grandmother, María Luisa

Acknowledgements

A story about love and smoking

Preamble

I used to be a heavy smoker.

II

When I mourned the passing of my beloved grandmother, I smoked day and night, one after the other. Two months after her death, I started coughing. I coughed day and night but I did not stop smoking.

By the fourth month, I had to quit smoking even though I did not want to. I would have a coughing fit every time I lit a cigarette.

Months later, I could not stand the smell of smoke anymore. There were two reasons: the first, my body was felt a strong nausea whenever a smoker walked alongside me; the second reason was that it reminded me of the saddest part of the mourning process.

That is how mental associations work. From then on, smoking would always be connected to the departure of my most loved one.

Years later, I realize that was one of the very last gifts she gave me: at all times, I have beautiful memories of her, but when I feel the need for a cigarette I remember the very sad mourning process and I cannot light up.

III

Crafting a Ph. D. dissertation is rewarding overall, but the process can be a frustrating and lonely one. In my case, sometimes it was extremely frustrating and full of anxiety.

Do not misunderstand me, I like research!! But I have the carpal tunnel syndrome, which basically means that repetitive movements with my fingers hurt. Typing with the keyboard is painful and using

a mouse takes me out of the office. I have to rely on computer control methods that can be slow, to say the least.

IV

The mental relief of tobacco is, on average, felt 7 seconds after a puff.

During the bad times with my thesis, I always felt the need for a cigarette. Then, a reminder of my grandmother's death was activated and the need for a cigarette evaporated.

V

I decided to train my mind to create positive memories and associations that could help to cope with frustration. I remembered how much she liked to watch me cooking something she had taught me and the great times we had at the kitchen together.

I would think as if I were "cooking" this thesis for her. I was able to dictate the thesis and bring positive associations without a cigarette.

I am a survey methodologist. I like to watch and analyze people's behaviour. I saw that two people that I admire professionally do not smoke: Willem and Melanie.

They practice sports regularly: speed skating and swimming. I followed their example and went back to the physical activity I like the most: dancing. I went for ballet, flamenco and tango.

I wanted to be a good role model for my dear niece, Danaé, providing her with an example of a person who how knows to address difficult issues without smoking.

VI

My boyfriend at that time was a "social smoker", he would smoke during weekend gatherings. After I quit I could not stand his smell when he did it. That was a serious issue...

He quit completely as well and went back to sports...

VII

I do not know if I have made a large contribution to survey methodology, but I tried my best. However, I do have a method for quitting smoking:

1) Rely on negative associations to do it e.g. a mourning process.

2) Find positive associations to cope with frustration to substitute a cigarette e.g. nice memories about my grandmother.

3) Rely on the example of the savvier ones to learn how to cool off e.g. Willem skates and Melanie swims.

4) Find a young person you want to show good traits in life e.g. my niece, Danaé.

5) Create positive externalities that keep you motivated: That guy who stopped smoking when I could not kiss him anymore is now my beloved husband, Andreu... And now we dance together.

Interlude

(Andreu, I just would not be able to finish this or any other silly project without you. I promise, I will not take on a Master in Statistics, a PhD and a job at the same time ever again ☺ But I do promise more adventures to keep us entertained. Thanks for your love, for helping me to give format to the thesis, for taking care of everything at home. But, I will not exempt you from reading it from cover to cover ☺)

VIII

I learnt several more personal lessons while doing the thesis, one of the most important from Willem, was to be precise.

And if I learnt it correctly I have to acknowledge that most of the credit for finishing this work goes to him. To his infinite patience with the many drafts of the articles. To his teaching of statistical science and survey methodology in a way that does not look difficult anymore.

He takes most of the credit because he has always been available to answer questions, to challenge my errors and to make me think about how to improve. For sharing his conviction on how to do science and being a model in life. And, of course, for always being precise.

Thank you, Willem Saris

IX

The list of close ones to thank is long both in my personal and in my professional life. Despite the challenges, doing a thesis is much easier when you have a reliable network. It starts with my co-supervisor Verónica Benet-Martínez. I appreciate her very useful comments each time a new article was ready and her support in the process. She is also a professional role model who, of course, does not smoke ☺.

At my research centre, I thank my close-knit colleagues who from 0 to 10, they score 11 in providing support. Anna, Melanie and Wiebke: I have learnt a lot from you, ladies. I am very happy to finish my outstanding deliverable and to get back to working hard alongside with you.

At the ESS Core Scientific Team, I have had the privilege of working with excellent survey methodologists and social scientists from whom I learnt a lot: Achim, Ana, Brina, Elena, Eric, Henk, Knut, Lizzy, Lorna, Sarah, Salima and Verena. Many thanks for asking how I was doing and providing support, articles and ideas through these years. My special thanks to Angelika, Geert, Ineke, Joost and Kirstine for their kindness and support.

At GESIS, I want to thank Angelika Scheuer and Beatrice Rammstedt for hosting me. During my visit, I wrote the literature review that later on went to each of the articles. There, I got closer to wonderful people: Dorotheé, Verena and Monika. My recognition to the inspiring ESS Translation Expert Task Group meetings where I work together with Allisu Schoua-Glusberg and Beth-Ellen Penell, two survey methodologists who have opened up the way for women in this discipline.

At UPF, I want to thank the support of my colleagues Andre (his willingness and motivation to discuss longitudinal analysis), Laur, Paolo and Teresa. I thank Professor Klaus Nagel whose kindness and nice conversations initially motivated me to start a PhD at UPF. At the ESRA Conferences where I presented my work, I'm thankful for the very useful comments from Eldad Davidov, Peter Schmidt and many other excellent researchers. My special thanks to Robin Motheral who took care of my awful English mistakes.

X

The last year of the PhD brought me to work more closely and to get to know better three people I deeply like and want to keep with, both professionally and personally: Brita Dorer, Malcolm Fairbrother and Rory Fitzgerald. Brita, thanks for listening the same story so many times, for inviting me to join so many interesting meetings, for our dinners. Malcolm and Rory, I think I rushed to finish to avoiding you asking the same question again: so when are you done? ☺ Thanks, it worked!!!

Then it comes my friends and family. A mis queridas flamencas: Cris, Laura, Eli y Bea ¡ya está, por fin! ¡Vamos a malear! To my dear friend Glòria, thanks for listening, always! To my dear Germans: Lisa and Lydia thanks for our debates, your moral support, Aperol spritz and for making life at UPF easier and even enjoyable. Camil, thanks for always listening my crazy ideas. Sebastián, gracias por creer en mí y por la música para acompañarme. En México, Glieb, File, Natalie, Lucero y Tania, mis personas favoritas ¡vamos a rockear! ya salió el ratón de la biblioteca. A la meva família catalana, especialment a Pilar i a l'avi Andreu. Gràcies pel vostre suport i per obrir les portes de Sant Vicenç on vaig acabar l'últim article. Gracias a mi madre, Guadalupe, a mi padre Artagnan, a mi hermana Brenda, a mi tía Estela, a mi primo Esteban por siempre confiar que puedo acabar lo que empiezo. Gracias a mi pequeña, querida y sabia sobrina Danaé, porque sus palabras son una fuente de ayuda para lxs adultxs confundidxs en su vida (She's 10 and rocks!).

XI

A tribute to Johann Sebastian Bach, his music sounds in the background through all my work.

Epilogue

My final words are of recognition to my English teacher Miss Lily. She taught me how to write in primary school. Thanks to her I learnt to command the language I use most at work. She did an excellent job, my current mistakes are only mine.

She was taken away too early, at age 56, by a lung cancer caused by a life of smoking.

Abstract

The present dissertation explores *language effects* in a comparative survey i.e. to what extent linguistic diversity affects equivalence in a comparative survey. This is done by studying three different dimensions on the challenges of designing a comparative multilingual survey: survey translation, linguistically diverse countries and bilingualism. Guidelines in survey translation do not link assessment criteria and measurement equivalence testing. I propose a systematic procedure to compare versions of a question in different languages before fieldwork which establishes that link. In linguistically diverse countries, survey instruments are translated into more than one language, equivalence is commonly assumed, not tested. I test for invariance distinguishing the response and cognitive processes to a survey question. Finally, I study measurement equivalence within an individual in two languages for political constructs (bilingualism), challenging current methodological approaches by bringing latent variable models.

In each dimension, findings aim to contribute to improving comparative survey methodology.

Resumen

Esta tesis explora los *efectos del lenguaje* en una encuesta comparativa: en qué medida la diversidad lingüística afecta la equivalencia de los datos mediante el estudio de tres dimensiones: la traducción de encuestas, países lingüísticamente diversos y,

bilingüismo. Las directrices actuales en la traducción de encuestas no vinculan los criterios de evaluación con un test de equivalencia. Se propone un procedimiento sistemático para comparar las versiones de una pregunta que establece dicho vínculo, en diferentes idiomas antes del trabajo de campo. En países lingüísticamente diversos, el cuestionario se traduce en más de un idioma. Se realiza un test de equivalencia que permite distinguir los procesos de respuesta de los cognitivos. Finalmente, se estudia la equivalencia de conceptos políticos en dos idiomas para un individuo (bilingüismo), proponiendo un enfoque metodológico de modelos de variables latentes.

Los hallazgos tienen por objeto contribuir a mejorar la metodología de encuestas en estudios comparativos.

Resum

Aquesta tesi explora els *efectes del llenguatge* en una enquesta comparativa: en quina mesura la diversitat lingüística afecta l'equivalència de les dades; mitjançant l'estudi de tres dimensions: la traducció d'enquestes, els països lingüísticament diversos i el bilingüisme. Les directrius actuals en la traducció d'enquestes no vinculen els criteris d'avaluació amb un test d'equivalència de mesures. Per tant, es proposa un procediment sistemàtic que estableix aquest vincle per comparar les versions d'una pregunta en diferents idiomes abans del començament del traball de camp. En països lingüísticament diversos, el qüestionari es tradueix en més d'un idioma. A la tesi, es realitza un test d'equivalència que permet

diferenciar els processos de resposta dels cognitius. Finalment, s'estudia l'equivalència de conceptes polítics pel mateix individu en els seus dos idiomes (bilingüisme), mitjançant l'aplicació de models amb variables latents.

Els resultats tenen per objectiu contribuir a millorar la metodologia d'enquestes en estudis comparatius.

Table of contents

Chapter 1

General Introduction

1. General Introduction

Large scale comparative cross-national surveys in the social and political sciences are measurement instruments of social attitudes, political opinions, preferences and behaviours across countries. They are multinational, multicultural and multilingual survey projects with the objective of making meaningful comparisons across populations of study. Researchers using comparative survey data to explain social and political phenomena rely on two assumptions: Firstly, that a relationship exists between concepts and unmeasured constructs (latent variables). Latent variables are then linked to manifest variables that are obtained from answers to the measurement instruments (survey items). Secondly, that this relationship is the same across countries.

In order to derive meaningful comparisons, a requirement must be met: measurement equivalence. The relationship between unmeasured constructs (latent variables), measurement instruments (survey items) and indicators (manifest variables) should be the same across countries.

Questionnaire design in a comparative survey should "maximize the comparability of survey questions across cultures and languages and reduce measurement error related to question design" (Janet A. Harkness et al., 2011, p. VI.3). Depending on the composition of cultural groups within participating countries, survey researchers decide in which language(s) the questionnaire is developed and in which languages translations are prepared. The consequences of

non-comparable measurement instruments across linguistic groups are that differences perceived as substantial may have their origin in differences in the way respondents react and understand measurement instruments.

The present dissertation is a compilation of three articles exploring *language effects* in a comparative survey i.e. to what extent linguistic diversity affects equivalence in a comparative survey. This is done by studying three different perspectives on the challenges presented in the design of a comparative multilingual survey: 1) survey translation, 2) linguistically diverse countries and 3) bilingualism.

The objective of this introduction is to provide a definition of equivalence in a multilingual comparative and to introduce its testing procedure. With this definition, the connection across the three articles is contextualized. The final section introduces the data sources used in the thesis.

1.1 Definition of equivalence

According to Scheuch (1993:113), equivalence as a requirement for comparability should be understood as whether questions are functionally equivalent for the purposes of the data analysis, i.e. in terms of statistical equivalence, not in terms of a common sense meaning of identical questions. This implies that the indicators obtained from survey items should represent, across groups, the same concepts they intend to measure (Mohler & Johnson, 2010).

Figure 1 below illustrates the definition of equivalence in a comparative survey with a measurement model, represented by Equation (1) (Bollen, 1989; Jöreskog & Van Thillo, 1973; Meredith, 1993; Saris, 1982b). In the figure, the parameters of the measurement model are the same for three linguistic groups. Their response function holds for statistical equivalence. In this case, the intercept and the slope of the response functions are the same. As a result, if two respondents belonging to two linguistic groups have the same opinion (same score, ξ_0, on the latent variable), their answer will be the same (y_0), the manifest variable.

In contrast, Figure 2 shows how linguistic diversity can affect equivalence. The intercept τ^g and/or the slope λ^g -or factor loading- of the response functions related to the answers to the survey items with the latent opinion are not the same for $Group\ 1$, $Group\ 2$ and $Group\ 3$. In this situation, the data cannot be compared across languages because, if groups have the same score in the latent variable, ξ_0, they would have given different answers. $Group\ 1$ would express its opinion in a more extreme way (y_3), while $Group\ 3$ would do the opposite (y_1), it would use moderate intervals for both low and high scores in the latent variable. The score of $Group\ 2$, y_2, would be somewhere between $Group\ 1$ and $Group\ 3$. Therefore, comparative survey research requires equivalence i.e. the same relationship between the latent and the observed scores for all groups.

$$y^g = \tau^g + \lambda^g \xi^g + \delta^g \text{forg} = \{1, 2, 3\} \qquad (1)$$

Figure 1. Measurement equivalence

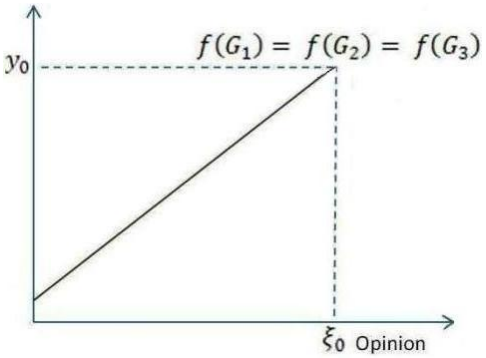
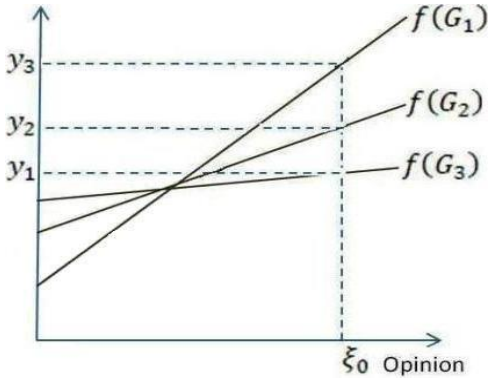


Figure 2. Measurement equivalence is not established



1.2 Testing for equivalence of survey data

Measurement equivalence or *measurement invariance* in comparative survey research is only confirmed by formally testing it. The test assesses whether the same measurement model holds across different groups or (sub)samples. Equation (1) shown above can be generalized into a model for p indicators i.e. manifest variables, represented by the $p \times 1$ vector $\boldsymbol{y}^g$ linearly related to m unmeasured constructs i.e. latent variables represented by the $m \times 1$ vector $\boldsymbol{\xi}^g$ for g linguistic groups $g = 1, \dots, n$ (Equation 3). The intercepts of the model are represented by the vector $\boldsymbol{\tau}^g$ $(p \times 1)$, $\boldsymbol{\Lambda}^g$ is a $p \times m$ matrix of factor loadings and $\boldsymbol{\delta}^g$ is a $(p \times 1)$ vector of disturbance terms. It is assumed that the expected value of the disturbance terms is zero, $E(\boldsymbol{\delta}^g) = 0$; that they are not correlated with the latent variables of the model $cov(\boldsymbol{\xi}^g, \boldsymbol{\delta}^{gT}) = 0$ and that the disturbance terms are independent (uncorrelated). (Bollen, 1989; cf. Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014; Jöreskog & Van Thillo, 1973; Meredith, 1993; Saris, 1982b; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000; Vandenberg, 2002).

$$\boldsymbol{y}^g = \boldsymbol{\tau}^g + \boldsymbol{\Lambda}^g \boldsymbol{\xi}^g + \boldsymbol{\delta}^g \quad (2)$$

From this measurement model, the following mean structure can be derived:

$$\boldsymbol{\mu}^g = \boldsymbol{\tau}^g + \boldsymbol{\Lambda}^g \boldsymbol{\kappa}^g \quad (3)$$

where $\boldsymbol{\mu}^g$ and $\boldsymbol{\kappa}^g$ are two vectors of observed and latent means of dimension $p \times 1$ and $m \times 1$ respectively. Finally, the following covariance structure can be derived:

$$\boldsymbol{\Sigma}^g = \boldsymbol{\Lambda}^g \boldsymbol{\Phi}^g \boldsymbol{\Lambda}^{g'} + \boldsymbol{\Theta}^g \quad (4)$$

where $\boldsymbol{\Phi}^g$ is the $m \times m$ variance-covariance matrix of the latent variables that may or may not have zeros in the off-diagonal elements. $\boldsymbol{\Theta}^g$ is the $p \times p$ variance-covariance matrix of $\boldsymbol{\delta}^g$.

The model, as represented by Equation (2) to Equation (4), has in each group, g, five matrices of parameters $\boldsymbol{\Lambda}, \boldsymbol{\tau}, \boldsymbol{\kappa}, \boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ to be estimated using survey data. For its identification, the scale of the latent variables has to be determined. This can be done by fixing one item's factor loading of each latent variable to one. In the same way, the parameters $\boldsymbol{\tau}^g$ and $\boldsymbol{\kappa}^g$ in Equation (3) cannot be estimated simultaneously, therefore, the intercept of the item which has already fixed its loading to one can be fixed to zero for each latent variable (Sörbom, 1982; Steenkamp & Baumgartner, 1998).

The test for measurement equivalence assesses whether the same measurement model holds across different groups or (sub)samples. It sequentially restricts parameters in Equation (2) to Equation (4) to be equal across groups. It is typically done in three steps, where each step is a prerequisite of the next one. In the first step, a *configural* model is fitted to check if the pattern of fixed and free loadings and disturbance terms is the same across groups (Horn &

McArdle, 1992). At this level of invariance, it is tested if (2)-(4) hold in each linguistic group without imposing any equality constraints in the parameters of the model.

In the second step, *metric invariance*, the configural model is restricted to one where the factor loadings of alike manifest variables are invariant across linguistic groups (5).

$$\boldsymbol{\Lambda}^1 = \boldsymbol{\Lambda}^2 = \cdots = \boldsymbol{\Lambda}^n \qquad (5)$$

When the model is not rejected, comparisons of relationships across groups can be made (Horn & McArdle, 1992).

The third step, *scalar invariance* implies that in addition to invariance in the factor loadings, intercepts of alike manifest variables are also restricted to be the same across groups (6).

$$\boldsymbol{\tau}^1 = \boldsymbol{\tau}^2 = \cdots = \boldsymbol{\tau}^n \qquad (6)$$

If the model is not rejected, comparisons of means can also be made across groups. Differences in the covariance of manifest variables can be attributable to, $\boldsymbol{\Phi}^g$, group differences in variances and covariaces of the latent variables and to, $\boldsymbol{\Theta}^g$, group differences in error variances.

In the second and third paper of this dissertation, in addition to testing for invariance in the measurement parameters, I explore

invariance between the structural parameters of the model by restricting $\boldsymbol{\kappa}$ and $\boldsymbol{\Phi}$ in addition to the restrictions in (6). Invariance of latent means implies imposing (7):

$$\boldsymbol{\kappa}^1 = \boldsymbol{\kappa}^2 = \cdots = \boldsymbol{\kappa}^n \qquad (7)$$

To test whether correlations among the latent variables are also invariant, two additional restrictions must be met. The first is *factor covariance* invariance, which is tested by restricting the off-diagonal elements of the $\boldsymbol{\Phi}$ matrix to be equal across groups the second is *factor variance* invariance which implies restricting the diagonal elements of the $\boldsymbol{\Phi}$ matrix (Steenkamp & Baumgartner, 1998).

In this dissertation, the models to test for measurement invariance are fit to sample data by maximum likelihood (ML) estimation by the means of multi-group structural equation modelling (MG-SEM), where the implied population covariance matrix, $\widehat{\boldsymbol{\Sigma}}^{\boldsymbol{g}}$ and mean vector, $\widehat{\boldsymbol{\mu}}^{\boldsymbol{g}}$ are estimated with the parameters of the model in a way that they are as similar as possible to the sample data formed by the sample covariance matrix, $\boldsymbol{S}^{\boldsymbol{g}}$ and the sample mean vector, $\bar{\boldsymbol{x}}^{\boldsymbol{g}}$ (Jöreskog & Van Thillo, 1973; Jöreskog, 1971; Sörbom, 1982).

Recent developments in Bayesian structural equation modelling (BSEM) make it possible to estimate the measurement invariance test under a Bayesian approach (for a review, see Davidov et al., 2014). In the Bayesian setting, in a methodology known as

approximate measurement invariance, equality constraints at the different invariance levels may not longer be fixed to a constant zero, but have a distribution with a location parameter at zero and a small variance, thus allowing for small differences in the constraint parameters (Muthén & Asparouhov, 2013).

Asparouhov and Muthén (2014) and Van De Schoot et al. (2013) used simulated data to show that although approximate measurement invariance can ease problems of badly fitting models, it can lead to underestimation of deviating parameters and overestimation of invariant ones. To solve this issue Asparouhov and Muthén (2014) proposed the *alignment* method, and optimization procedure which does not assume exact measurement invariance to estimate latent means and factor variances, but estimates them while at the same time optimizes the measurement invariant pattern i.e finds a solution with the minimum number of deviating parameters with the largest differences across groups.

Although these developments will potentially ease estimation of models that test for measurement invariance, and applications of these estimation procedures using comparative survey data are growing (for applications to measurement invariance of human values scales, see Cieciuch, Davidov, Schmidt, Algesheimer, & Schwartz, 2014; Zercher, Schmidt, Cieciuch, & Davidov, 2015) further work needs to be done to explore whether or not and underwhat conditions results are different to the classical frequentist approach. For instance, in the articles in this dissertation, the

alignment method cannot be implemented, because a current limitation of the procedure is that observed variables should only load on one factor (models without cross-loadings).

1.3 The link across the articles

a) Survey translation

When testing for invariance, several studies have identified translation decisions as a source of non-equivalence (Brislin, 1970; Davidov & De Beuckelaer, 2010; Hambleton, Merenda, & Spielberger, 2005; Janet A. Harkness, Villar, & Edwards, 2010; Mallinckrodt & Wang, 2004; Oberski, Saris, & Hagenaars, 2010; Van de Vijver & Leung, 1997; Van de Vijver & Tanzer, 2004). Unfortunately, translation deviations were detected after data was already collected. Therefore, survey translation requires developing procedures that are oriented to enhance equivalence before data collection.

Survey translation has developed best practice procedures for translating a questionnaire to get *equivalent* survey instruments in multilingual contexts. Translation guidelines suggest that a good translation aiming at functional equivalence would avoid deliberately changing semantic components other than those necessary because of language differences. This means that a translation should keep the same concepts across languages; preserve the item structure and maintain the intended psychometric properties (Harkness, Pennell, and Schoua-Glusberg 2004;

Harkness, Villar, Edwards 2010; Harkness 2003). However, guidelines do not suggest systematic approaches to assess whether a resulting translation is equivalent.

Current practices in translation quality assessment do not have a direct link with a measurement invariance model. Most of the assessment procedures require the judgment of evaluators. Judgments may be subjective or evaluators may focus on only one set of elements to assess an item. The final decision about the appropriateness of a translation should not rely on one (or a team of) expert(s), but on model-based evidence.

This shortage of methods to empirically test questionnaires motivated the research question in the first article: *How to detect deviations, in terms of measurement equivalence, of a survey instrument in different languages before it is administered to respondents?*

The method proposed in this dissertation consists in comparing the features of source and target survey items in the same coding scheme. The characteristics that are compared determine the form of the items and are predictors of its measurement quality. I propose doing this using the coding scheme in the Survey Quality Predictor (SQP) Software (Saris et al., 2011). The software is based on a large collection of multi-trait multi-method (MTMM) experiments. With the MTMM approach, the measurement quality is estimated as the square product of the standardized reliability and validity

coefficients. This model is an equivalent model to the general one shown in Equations (2)-(4). The measurement quality indicates the strength of the relationship between the observed variable and the latent variable of interest.

The procedure was implemented in Round 5 to Round 7 in the European Social Survey. The results show that there are avoidable differences in the formulation of the questions due to translation. If they were not prevented, they would have potentially impacted measurement equivalence.

b) Equivalence in multilingual countries

Many countries are not linguistically homogeneous as various languages are spoken at the regional level. In the case of countries where instruments in more than one language were administered, it cannot be assumed that the data is statistically equivalent without a test. As language and culture are strongly interrelated (Cohen, 2009; Sam & Berry, 2010), multilingual countries are culturally heterogeneous. Cultural orientations can influence the connotations and appropriateness of a survey question (Van de Vijver & Leung, 1997; Van de Vijver & Tanzer, 2004). Moreover, cultural specifics can prompt respondents' reaction to elements of the measurement method. For instance, evidence suggests significant differences across cultures in the use of extreme and mid-point categories of response scales (Harzing, 2006).

Given that the form of the questions can be checked and the type of differences controlled (Article 1/Chapter 2), it is important to know if the type of differences resulting in non-equivalent data are based on cultural differences. One type of cultural difference that impacts comparability is related to the way in which the concepts asked in the survey are understood across groups. A second type refers to differences in the way respondents of different cultural background react to the measurement method i.e. the combination of characteristics that define the formulation and administration of the request, such as the response scale, the mode of data collection, the use of showcards or visual aids, the translation procedure, the selection and assignment of languages, the introduction, the additional explanations, among others (Saris & Gallhofer, 2014).

In Article 2 (Chapter 3), my coauthor and I challenge a current practice to test for invariance by defining the groups at the country level without testing first across linguistic groups. We use a model that allows separating the respondent's reaction to the measurement method, the *response process*, from a true difference in the interpretation of the meaning of concepts, the *cognitive process* (Saris & Gallhofer, 2014). This model extends over the classical parameterization to test for measurement invariance working on a common criticism that has been referred to as *susceptibility* i.e. to what extent the procedure is sensitive to artifacts in the response process (Butts, Vandenberg, & Williams, 2006; Byrne & Watkins, 2003; Marsh & Byrne, 1993). Recent developments in the measurement invariance literature suggest that when non-invariant

patterns are found at the measurement level, it should be assessed to what extent they impact substantive conclusions i.e. to what extent non-equivalence of measurement instruments biases the estimates of parameters of interest (Meuleman, 2012; Oberski, 2014).

Equation (8)-(10) show an example of how the model is extended in this article, as such it is not identified, but serves as an illustration in this introduction. Equation (8) and (9) represent a response process where y_1^g and y_2^g are two manifest variables obtained by asking the same question in two different forms. They are linearly related to the latent variable η_1^g. Method factors for each language are η_2^g and η_3^g. The cognitive process is represented by Equation (10) having α_1^g as the intercept and γ_{11}^g as the regression coefficient of η_1^g on ξ_1^g.

$$y_1^g = \tau_1^g + \lambda_{11}^g \eta_1^g + \lambda_{12}^g \eta_2^g + \delta_1^g \qquad (8)$$

$$y_2^g = \tau_2^g + \lambda_{21}^g \eta_1^g + \lambda_{23}^g \eta_3^g + \delta_2^g \qquad (9)$$

$$\eta_1^g = \alpha_1^g + \gamma_{11}^g \xi_1^g + \zeta_1^g \qquad (10)$$

With this model, the elements of the variance-covariance matrix of the model, $\boldsymbol{\Sigma}^g = \boldsymbol{\Lambda}^g \boldsymbol{\Phi}^g \boldsymbol{\Lambda}^{g'} + \boldsymbol{\Theta}^g$, are decomposed. The $\boldsymbol{\Lambda}^g$ and $\boldsymbol{\Theta}^g$ elements are a combination of the response and the cognitive parameters.

In the paper, it is argued that invariance should not necessarily be required in the parameters of Equations (8) and (9), that represent respondents' reactions to the formulation of the measurement instrument in two languages. Invariance is sufficiently required in the structural parameters representing the way respondents interpret survey items (10). Separating the respondent's reaction to the measurement method, the *response process*, from a true difference in the interpretation of the meaning of concepts, the *cognitive process,* is important because if differences in parameters across groups causing non-invariance have their origin at the response process, invariance can be established by correction for measurement error.

The article uses data from the Round 2 of the European Social Survey analyzing equivalence within country in four multilingual countries: Belgium, Estonia, Switzerland and Ukraine. Results show that non-invariance was most common in the response process rather than at the cognitive process.

Recent developments in measurement invariance include individual or group level predictors to account for cross-cultural differences and explain non-invariance using hirarchichal (multilevel) models (Davidov, Dulmer, Schluter, Schmidt, & Meuleman, 2012), however multilevel modelling requires that a large number of groups are compared in the tests: above 50 in a frequentist estiamation (Meuleman & Billiet, 2009) and about 20 in a Bayesian one (J. J. C. M. Hox, van de Schoot, & Matthijsse, 2012). This

approach is not useful to explain noninvariance due to the administration of two questionnaires in different languages within a country. Therefore, the article in chapter 3 offers an alternative to account for cultural differences when the number of groups is small.

c) Is it possible to have two opinions? Bilingualism in survey research

Sociocultural psychologists and psycholinguistics have studied since the mid-sixties whether for psychological constructs, the opinion of individuals in one or another language is the same (cf. S. X. Chen & Bond, 2010; Ji, Zhang, & Nisbett, 2004). The third article in this dissertation explores if that is the case for survey items measuring political attitudes. The study of language effects has a long tradition in psychological instruments but it has arrived with delay to survey methodology. Given that, in Article 2 (Chapter 3), cognitive invariance was generally established, and that it is possible to correct for non-invariance in the response process, in Article 3 (Chapter 4), I explore whether the correlation of the same latent constructs in different languages is equal to one.

Evidence from psychology shows that the language of an interview can activate cultural orientations driving individuals' responses (S. X. Chen & Bond, 2010; Luna, Ringberg, & Peracchio, 2008; Ramírez-Esparza et al., 2006; Schwartz et al., 2014). Language is a strong cultural carrier (Cohen, 2009) and bilingual individuals tend to live in mixed cultural environments. Cultural orientations may influence thoughts, cognitions and behaviour (Oyserman & Lee,

2008) and this in turn may affect the way respondents interpret and answer survey questions.

Evaluating whether language effects are present in survey items is still relevant for three reasons. Firstly, measurement invariance for the same individual in two languages has seldom been established prior to test for language effects. Secondly, language effects are commonly assessed using a test for mean differences in composite scores. When differences in observed means have not been found significant, the conclusion has been that language effects are negligible. However, by comparing mean scores, it is not tested if the conceptual associations that individuals retrieved when they use one language or the other are the same.

I propose testing for language effects using the application of a LISREL model (Jöreskog & Van Thillo, 1973) which assumes linear relationships between indicators (observed variables) and unmeasured constructs (latent variables). With this linear model (Equation (2) - Equation (5), in the first step, I test if the relationship across indicators and latent variables is the same for an individual in two languages. This is the test for measurement invariance.

Once it is established that the measurement model is equivalent, I am able to test structural relationships of latent variables in two languages. I tested if two latent variables represent the same variable of interest by testing if its correlation is equal to one

(Jöreskog, 1971; Saris, 1982a, 1982b). I show that a test where latent (or observed) mean differences are not significant does not rule out the possibility of language effects. It indicates that the distribution of the variable in two languages is the same, as the location parameter is the same, but respondents can still have different conceptual associations in each language. Furthermore, language effects are interpreted as distinct associations that individuals make in each language. They have implications for fieldwork procedures which are discussed in the article.

1.4 Data sources

a) European Social Survey

Articles compiled in Chapter 2 and Chapter 3 used data from the European Social Survey (ESS). The ESS is an academically driven survey conduted each two years in about 25 European countries including Russia and Israel. It is administered in a face-to-face closed interview to a probabilistic sample of respondents in each participating country. The ESS has the objective to comparatively measure social and political attitudes, opinions and behaviours in the countries where it is fielded (European Social Survey, 2016).

The ESS is a representative survey of "all persons aged 15 and over resident within private households in each country, regardless of their nationality, citizenship or language" (European Social Survey, 2015b, p. 23). The effective sample size is 800 interviews for

countries with less than two million persons in the target population or 1,500 interviews otherwise with a target response rate of 70%.

In Chapter 2, the unit of analysis was the survey questions. The five-step procedure to compare the codes the source questionnaire and translated language versions was applied in a sample of questions from Round 5 (fielded in 2010), Round 6 (2012) and Round 7 (2014) of the ESS as a last step of quality control in the translation procedure. A total of 102 questions have been evaluated. 34 questions include the topics "Trust in criminal justice", "Attitudes towards immigration", "Personal and social well-being", "Democracy" and "Political efficacy" and 68 are repetitions of them with a variation in the measurement properties designed for experimental purposes[1]. Translations are obtained when at least 5% of the country's population is native speaker of a language.

In each round, the ESS Core Scientfic Team (CST) designs a number experiments in a supplementary questionnaire in which questions with variations in the measurement method are repeated to the same respondents (within-subject design). In Chapter 3, I use data from the European Social Survey Round 2 (European Social Survey, 2005) because in that round, repetititons were administered for the concept of political trust. This made it possible to have a multiple indicators model. Countries selected were Belgium, Estonia, Switzerland and Ukraine because the proportion of

[1]The formulation of the items in the main and supplementary questionnaires as designed in the English Source version is available at http://europeansocialsurvey.org

respondents in minority linguistic groups is at least 25% covering a diverse range of languages: French, Dutch, German, Estonian, Russian and Ukrainian.

b) The Longitudinal Internet Studies for the Social sciences Immigrant Panel

For Chapter 4, I applied to the Measurement and Experimentation in the Social Sciences (MESS) Immigrant Panel administered by CentERdata at Tilburg University, The Netherlands. The Longitudinal Internet Studies for the Social sciences (LISS) Immigrant panel was a probability-based online project in which researchers of various social and political fields could submit proposals for fieldwork at no cost.

It consisted of around 1600 households/2400 respondents with about 1100 households/1700 respondents of non-Dutch origin (CentERData, 2010). Respondents were recruited based on stratified sampling using the population registry as sampling frame. Participants had foreign backgrounds of four major migration groups in the Netherlands (first and second generations of western and non-western origin). They were provided with internet and a laptop to answer monthly surveys and received an economic incentive for each completed questionnaire.

The study for this dissertaton spanned over two waves between April and June, 2013. In Wave 1, the objective was to select the languages in which translations would be obtained to test for language effects in a within-subject design in Wave 2. Wave 1

included 989 bilingual participants. They mentioned 74 languages as their native tongues. I selected the five languages in which respondents had the highest self-reported proficiency and the group was of at least 30 individuals: Arabic, English, German, Papiamento and Turkish. The source questionnaire was developed simultaneously in Dutch and English. I coordinated the translation process. Translations into the other four target languages were done according to a committee approach with two independent translators and an adjudicator that decided over differences between translations, questions were pretested with at least one person in each language. This approach was based on the TRAPD (Translation, Review, Adjudication, Pretesting and Documentation) procedure which is the state-of-the-art procedure for translating survey questionnaires (Janet A. Harkness, 2003).

In the second wave, the questionnaire was presented to 308 bilingual panel members, and it was fully completed by 255 respondents (83%). Due to the small number of individuals per language, the analysis was done by linguistic group.

Chapter 2
A procedure to prevent differences intranslated survey items using Survey Quality Predictor

2. A procedure to prevent differences intranslated survey items using Survey Quality Predictor[2]

Abstract

Survey translation has developed best practice procedures to translate survey instruments aiming that the same stimuli and measurement properties should be provided across languages. Monitoring the formal structure of translated questionnaires in cross-sectional surveys is challenging. Current procedures in translation assessment do not link the quality of the translation with a formal test of measurement equivalence. In this article we present a procedure to prevent differences in the form of translated survey instruments by comparing with a common coding scheme the features that determine the measurement quality. The coding scheme is included in the Survey Quality Predictorsoftware (SQP). We present the results of the implementation of this procedure in Round 5, 6 and 7 of the European Social Survey.

2.1 Introduction

This article presents a procedure using the coding scheme of the Survey Quality Predictor software (SQP) (Saris et al., 2011) to prevent differences in the form and measurement properties of

[2] This paper received the 2014 Janet Harkness Student Paper Award (Honorable Mention) awarded by WAPOR/AAPOR at the 67th WAPOR Conference in Nice, France. An adapted version of this chapter is accepted for publication in the book: *Advances in Comparative Survey Methodology (forthcoming 2017),* Johnson, T., Penell, B., Stoop, I., Dorer, B., (Editors), Hoboken: Wiley& Sons.

survey questions during the translation process. We propose that deviations in translations that impact cross cultural equivalence can be detected if the characteristics of questions' form and the measurement properties are coded and the codes are compared across languages. It is proposed that this comparison should be made using the coding scheme of SQP software3 because the codes are independent of the languages. The coding scheme in SQP facilitates the exchange of information about item characteristics and measurement properties that should remain constant for translation teams and questionnaire designers. Once the characteristics are coded, these elements are the ones that need to be compared in order to detect deviations across language versions.

Survey translation has developed best practice procedures for translating survey instruments, which aim to provide the same stimuli and measurement properties across languages. We argue that current procedures in translation assessment do not help to check, in a systematic way, if translations are fulfilling best practices. This is mainly due to three flaws: 1) current procedures in translation assessment do not link the quality of the translation with a definition of cross-cultural equivalence. 2) Monitoring the formal structure of translated questionnaires in cross-sectional surveys is challenging because one cannot be familiar with all languages participating in a cross-cultural project. 3) Without an inventory of the elements that should remain constant across languages, it is very

[3] Available at sqp.upf.edu

difficult to check, in a systematic way, that the characteristics of the questions are the same across language versions.

Survey research in questionnaire design has studied how different features of question wording, layout of questionnaires and aid material, and answer scales affect respondents. When a questionnaire is translated, the translation team chooses from different wording options to remain equivalent with the source text. They also take decisions about the layout of show cards or of questions. There is very little research on objective criteria to decide among different translation options (cf. Behr, 2009). Therefore, translation assessment has remained a very subjective exercise. Harkness, Villar, et al. (2010) advised that problem solving in survey translation should start with the definition of the translation unit (the survey item), its goal (match intended meaning and intended measurement properties) and its audience (respondents) rather than focusing the discussion on the level of words.

In this context, we suggest that the criteria for deciding among translation options should be to preserve the item characteristics in both source and target versions (as long as the structure of the target language allows it). Those item characteristics have been defined by the tradition of questionnaire design in survey research and are included in the coding scheme of SQP program.

In this way, the procedure presented in this article which we call SQP Coding helps to provide criteria to monitor the quality of a

translation based on criteria that link the characteristics of the translated question with a definition of functional equivalence. After this introduction, Section 2.2 presents a framework of functional equivalence in cross-cultural research. Section 2.3 reviews the literature in survey translation and translation quality assessment to argue that current procedures do not link translation evaluation to a definition of equivalence. It is also argued that most procedures rely on subjective judgements and do not systematically monitor if key measurement elements in the translations remain the same across languages.

Section 2.4 defines the formal characteristics of a survey item (domain, concept, response scale, polarity, labelling, symmetry, balance of the request, introduction, instructions, linguistic complexity, layout of the question and of the aid materials, et cetera). It is argued that functional equivalent questions designed specifically for comparative research should keep these characteristics fixed across language versions to the extent the language structure allows it. The section describes the coding scheme in the Survey Quality Predictor software to collect and compare information in a systematic way about a comprehensive number of survey item characteristics.

Section 2.5 explains in detail the procedure “SQP Coding” for systematically comparing the formal characteristics of a source and a translated question as a means to detect deviations in the translation and layout of the questionnaire before it is administered

to respondents. Section 2.6 summarizes the findings of the implementation of SQP Coding in 102 questions of Round 5 (2010), Round 6 (2012) and Round 7 (2014) of the European Social Survey (ESS) in more than 24 languages. Section 2.7 ends the chapter discussing the findings and pointing out future research lines for cross-cultural survey translation.

2.2 Equivalence in survey translation

Survey methodology has made a distinction between *comparing national surveys* and implementing *comparative surveys from design* (Janet A. Harkness, Braun, et al., 2010). The difference is that a *comparison of national surveys* involves comparing surveys designed for a specific country, whereas *comparative surveys from design* are surveys thought to implement the same procedures and to have the same characteristics with the idea of matchings findings in each population of study. In the second type of survey, it is assumed that by keeping survey features the same -to the maximum possible extent- the data would be comparable. An example of a comparative survey from design is the Programme for International Student Assessment (PISA), which surveys how well 15-year-old students are prepared regardless of the curriculum taught in different schools across participating countries (Fleischman, Hopstock, Pelczar, & Shelley, 2010). The tests are designed in such a way that they aim to reflect the differences in the analytical tools of the students and not the cultural context in which education is embedded.

Therefore, the objective of *comparative survey research* is that the measurement instruments (questions) administered across populations are in fact *comparable*. According to Scheuch (1993:113), *comparability* should not be understood in terms of "whether [questions] are identical or equivalent in the commonsense meaning, but whether they are functionally equivalent for the purposes of analysis". For Mohler and Johnson (2010:23) Scheuch's definition implies that "functionally equivalent indicators are revealed in analysis, they cannot be judged on the basis of face value similarity. (…) they should behave in a similar manner in statistical analysis". This implies that the responses obtained should use the same measurement instruments and should represent the same concepts they intend to measure across groups.

For a survey questionnaire, equivalence has two conditions: 1) respondents should understand the survey questions in the same way across languages, i.e. they should understand the same concepts of interest asked via questions and 2) they should express themselves in the same way, i.e. the same opinion should correspond to the same observation answer across cultural/linguistic groups (Saris & Gallhofer, 2014; Saris, 1988).

Survey translation has developed best practice procedures to get *functionally equivalent* survey instruments in multilingual contexts. Procedures bring together the state of the art in translation studies and the particular needs of survey research. In translation studies, the concept of *functional equivalence* has already been discussed

for a long time. It requires that the message embedded in a text is received by the receptor *in the same way* as it would be received in the source language (Nida, 1964).

Translation guidelines suggest that a good translation aiming at functional equivalence would avoid deliberately changing semantic components other than those necessary because of language differences (Janet A. Harkness et al., 2004; Janet A. Harkness, Villar, et al., 2010; Janet A. Harkness, 2003). This implies that although a literal (word-by-word) translation is not required, questions should maintain the same concepts of interest across languages, preserve the item characteristics and maintain the intended psychometric properties. However, guidelines do not suggest how to formally test that a resulting translation is equivalent. In practice, it is very difficult to empirically check if the requirements set by translation guidelines –to maintain the intended psychometric properties and to keep concepts the same— are achieved because one cannot understand all languages. As Smith (2004:446) points out "perhaps no aspect of cross-national survey research has been less subjected to systematic, empirical investigation than translation."

With some exceptions in the context of the translation of psychological instruments and educational testing, there is little research on how to statistically assess cross-cultural instruments before they are administered to respondents (Brislin, 1970, 1976; Dean, Caspar, McAvinchey, Reed, & Quiroz, 2007; Hui & Triandis,

1985). Statistical procedures for checking the equivalence of measurement instruments across countries are improving and becoming more sophisticated but they are mostly helpful for detecting flaws once data is already collected. (Byrne & Van De Vijver, 2010; Meredith, 1993; Muthén & Asparouhov, 2013; Saris & Gallhofer, 2014; Van De Schoot, Schmidt, De Beuckelaer, Lek, & Zondervan-Zwijnenburg, 2015; Vandenberg & Lance, 2000).

This shortage of methods for empirically comparing questionnaires motivated the development of the approach introduced in this chapter: *how to detect differences affecting functional equivalence in a multilingual survey instrument before it is administered to respondents?*

2.3 Definition of measurement equivalence

There is consensus among survey methodologists that *measurement equivalence*– or measurement invariance- is a prerequisite for deriving substantive conclusions from data collected in diverse populations. It should not be assumed but tested that survey instruments measure the same constructs in exactly the same way across groups (Saris & Gallhofer, 2014; Van De Schoot et al., 2015; Vandenberg & Lance, 2000; Vandenberg, 2002).

The definition of measurement equivalence is illustrated below. In Figure 3, the intercept and/or slope of the response functions that relate the responsesto the latent opinion are not the same for

$Group1$, $Group2$ and $Group\ 3$. In this situation, the data cannot be compared across languages because if the groups hold the same score in the latent variable, ξ_0, they would have given different answers. $Group\ 1$ would express its opinion in a more extreme way (y_3), while $Group\ 3$ would do the opposite (y_1).It would use moderate intervals for both low and high scores in the latent variable. The answer of $Group\ 2$, y_2, would be somewhere in between $Group1$ and $Group\ 3$. In contrast, Figure 4 shows how the response function looks when it is statistically equivalent for the three groups. In this case, the intercept and the slope of the response functions are the same. As a result, if two respondents belonging to two linguistic groups have the same opinion (same score, ξ_0, in the latent variable), there should be correspondence in their observed answers (same score $y_0$in the manifest variables).

Figure 3. Measurement equivalence is not established

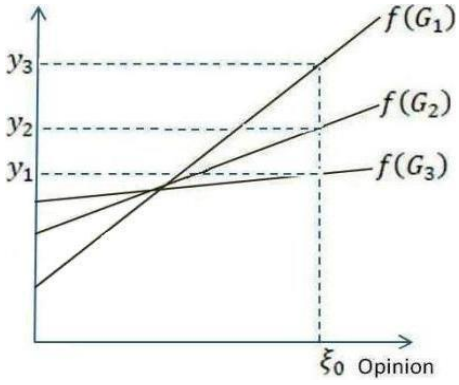
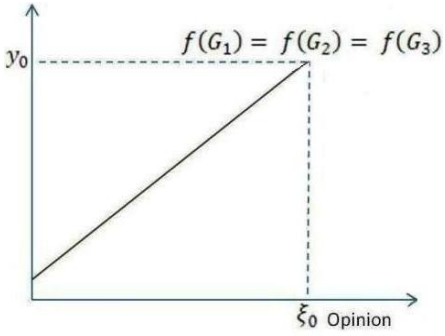


Figure 4. Measurement equivalence



Equivalence in cross-cultural survey research is confirmed by formally testing it. A typical statistical test for equivalence (*test for measurement invariance)* has three steps. At the third step, scalar invariance, the response functions are restricted to be the same across groups as is illustrated in Figure 4. If the test is not rejected, comparisons of means and relationships can be done across groups.

When scalar invariance is rejected, the responses are affected by *item bias* and/or *method bias*. Van de Vijver and Tanzer (2004:127) suggested that the most frequent causes of item bias are "item translation, ambiguities in the original item, low familiarity/appropriateness of the item content in certain cultures, or influence of cultural specifics such as nuisance factors or connotations associated with the item wording".

Method bias occurs when the observed answers are affected by a factor that is independent of the construct of interest and related to the characteristics of the measurement instrument, e.g. the response scale, layout of batteries, et cetera (Alwin, 2007; J. A. Krosnick & Fabrigar, 1997; Saris & Gallhofer, 2014; Van Herk, Poortinga, & Verhallen, 2004). This has been confirmed by research identifying translation decisions as a source of non-equivalence in assessments of survey data (Davidov & De Beuckelaer, 2010; Hambleton et al., 2005; Janet A. Harkness, Villar, et al., 2010; Mallinckrodt & Wang, 2004; Oberski, Saris, & Hagenaars, 2007; Van de Vijver & Leung, 1997; Villar, 2009). Unfortunately, the impact of translation decisions in data equivelence was detected once data was collected

and survey organisations had already spent a lot of resources on data collection.

2.4 Cross-cultural survey translation and translation assessment

The most widely used approach in questionnaire design for multiple cultures is frequently referred as the "Ask-the-same-question" model (Janet A. Harkness, 2003). In this model, a measurement instrument is designed in one or two languages called *source* (source language, source questionnaire, source item, source instrument, et cetera) and is adopted and exported to other settings called *target languages* via questionnaire translation and layout formatting.

In this section, we review current practices in survey translation and translation quality assessment. Translation assessment requires the judgement of evaluators. Judgements may be subjective or may focus on just one set of elements. Even if the translation procedures are under strict guidance, each translation has specific cultural elements, grammatical characteristics and a subjective inherent component. The final decision about the appropriateness of a translation relies on one person or a team of experts, but not on model-based evidence (Saris, 2012). In addition, most procedures do not have a direct link to a framework of cross-cultural equivalence.

Best practices recommend that translation should be integrated with questionnaire design rather than implemented when the questionnaire in the source language is finished (Erkut, Alarcón, Coll, Tropp, & García, 1999; Janet A. Harkness, Villar, et al., 2010). Harkness (2003) suggested a procedure called TRAPD an acronym for T: *Translation*, R: *Review*, A: *Adjudication*, P: *Pre-testing* and D: *Documentation* in which translations are done using a team -or committee- approach in a multistep process where different members provide expertise to arrive at a final translation. For instance, in the ESS and in the Survey of Health, Ageing and Retirement in Europe (SHARE), the translation committee is formed by a team including two translators, one reviewer and one adjudicator (European Social Survey, 2014; Janet A. Harkness, 2005, p. 24). Its members combine survey knowledge and linguistic expertise. The two translators make parallel translations from the source version to the country's language. The reviewer assesses the translation and the adjudicator is responsible for making decisions on the different translation options. The whole process is documented and the translated questionnaire is pre-tested.

It is claimed that in the TRAPD procedure quality monitoring is part of the process as changes are approved by a team and documented at each step (Janet A. Harkness, 2003). However, the way adjustments are decided remains a subjective exercise. Willis et al. (2010) evaluated the TRAPD procedure in five large scale cross cultural projects showing that its success pretty much depended on the team members' familiarity (translators, reviewers, adjudicators,

cognitive interviewers) with the purpose of the translation. A second quality monitoring problem in the TRAPD is that the documentation step produces a large amount of information which lacks systematic analysis and which may be burdensome for the average user (Mohler, Pennell, & Hubbard, 2008; Mohler & Uher, 2003).

A complementary procedure to the TRAPD is to conduct *advance translation* (Dorer, 2011; Janet A. Harkness, 1998). In this approach, a survey questionnaire is translated using the TRAPD approach during the questionnaire design stage into more than one language to foresee potential difficulties. For instance, the Eurobarometer survey -which also translates their instruments using a committee approach- and the Programme for International Student Assessment (PISA) design their source questionnaires in both English and French and they are taken jointly to produce target versions in other languages in a committee approach (Eurobarometer, 2012; OECD, 2012).

If advance translation shows challenges in the formulation of the source text, it could be modified to convey that. One limitation of this approach in surveys with more participating languages than the ones used in advanced translation is that when a problem is detected in one language and the source questionnaire is changed, another problem could appear in another language that did not participate in the advanced translation. Questionnaire designers would remain unaware of this second problem.

2.5 Translation assessment

Current procedures for translation assessment include back translation, translation verification and pretesting. Ideally, methods to assess a translation in survey research should evaluate whether the target text kept the semantic content and the psychometric properties determined by the item characteristics the same across languages (Janet A. Harkness, 2003).

A very common method to evaluate translated measurement instruments is *back translation* (Brislin, 1970, 1976). In this procedure the target questionnaire is translated back into the source language. Differences between the two texts are rendered as potential translation problems. This approach is necessary in order to make it possible for a translation to be understood by different members involved in the survey design process. However, as an assessment method, it is not exempt from limitations. The main criticism of this approach is that the target text is not evaluated, only a version of it in the source language. Other criticisms are that translators may use words that make a translation closer to the source but are incoherent in the target language because their own performance is evaluated taking back translation as a standard rather than as a tool. Deviations may not relate to the translations but to unmatched linguistic structures in both languages (Janet A. Harkness, 2003; Pan & De La Puente, 2005).

A recently applied method for survey translation is the outsourcing of semantic verification of target instruments. This procedure has been called *translation verification* or *semantic quality control*. It is used in projects such as the PISA, the Trends in International Mathematics and Science Study (TIMMS) and the ESS (Dorer, 2013b; Fleischman et al., 2010; Mullis, Martin, Foy, & Arora, 2012; OECD, 2012). An external provider verifies a questionnaire or a selection of items in all participating languages based on categories for potential interventions -among them: additional information, missing information, grammar/syntax, and consistency- to recommend changes in a translation when they are considered necessary. Verifiers give suggestions for improving countries' translations and the overall comparability of data; they also check compliance with annotations provided in the source questionnaire to produce more precise translations.

It has been found that there are differences between the verifiers' scope across languages (Dorer, 2013a). Some of them were more inclined to stylistic interventions while others were more inclined to verify content of the measurement instruments. Another potential problem is that the usefulness of the interventions is related to the verifiers' knowledge of each country's context. For example, in addition to regional differences in Russia, Russian is a minority language shared by several countries in Europe. Speaking populations use different forms and words to assign meaning in Russia, Israel, Estonia, Latvia and Lithuania. Verifiers need to be

familiar with several but at the same time proper usages and forms of the language in each country.

A procedure addressed to directly test the equivalence of survey instruments is both qualitative and quantitative *pretesting*. Pilot studies are pre-testing studies which require gathering large amounts of data, which is a costly and time-consuming procedure unaffordable for many cross-cultural surveys. For instance, the PISA conducts a pilot study with an average of over 200 student responses in most participating countries in each round. The data is used to eliminate items that are not statistically equivalent across countries, using common differential item functioning (DIF) and item response theory (IRT) techniques (PISA 2010). Pretesting in many projects mostly means to administer the questionnaire to a small group of respondents before starting fieldwork.

Another type of pretesting studies extensively used in psychology, which also require large amounts of data, are split ballot experiments with bilinguals individuals to test the equivalence of items in two language (John, Goldberg, & Angleitner, 1984). In these experiments, each random group answers the questionnaire in a different language. The reliability of the instrument is assessed considering the differences between the two groups (Mallinckrodt & Wang, 2004; Segalowitz, Hulstijn, Kroll, & de Groot, 2005).

Benet-Martínez and John (1998) and John et al. (1984) used multi-trait multi-method (MTMM) experiments to assess cross-language

validity in personality measures. Repetitions of the same traits in different languages were answered by bilinguals, making it possible to estimate the effects of language differentiating it from other sources of measurement error. However, this approach has limitations, research has shown that bilinguals do not use language in the same way as monolinguals do. Bilinguals may switch their cultural frame of reference depending on the language they use to answer (Blais & Gidengil, 1993; Bond & Yang, 1982; S. X. Chen & Bond, 2010; Ellis, 1992; Hong, Morris, Chiu, & Benet-Martínez, 2000; Yang & Bond, 1980).

Responses to attitudes and personality traits have varied depending on how integrated or conflicted the different cultural schemasare in bilinguals (Benet-Martínez & Haritatos, 2005). Thus, bilinguals seemed to follow different response patterns in each language depending on how integrated both cultures were in their own identities (Ramírez-Esparza et al., 2006). These results have decreased the validity of experiments using bilingual individuals as a means to test equivalence of translated instruments.

A very common pretesting method for multilingual instruments is *cognitive interviewing*. It helps to detect if concepts are understood similarly with a small amount of data. In its typical design, "think-aloud" and probing questions are used in a face-to-face interview to get information about item comprehension and response formulation (Beatty & Willis, 2007; Fitzgerald, Widdop, Gray, &

Collins, 2011; Pan & De La Puente, 2005; Pan, Landreth, Hinsdale, Park, & Schoua-Glusberg, 2007; Willis, 2004).

Pan et al. (2007) showed that respondents participating in cognitive interviewing in four different languages had in each group specific patterns of linguistic behaviour and communicative style. This meant that for the same probing questions, respondents differed in the way they answered the cognitive interview and not in the way they understood the survey item that was evaluated. Other criticisms are regarding the large effects of interviewers (Beatty & Willis, 2007; Goerman & Caspar, 2010), the thresholds for problem acceptance and the reliability of respondents in problem detection (Conrad & Blair, 2004).

Dean et al. (2007) have been pioneers in suggesting a coding tool for pre-testing cross-cultural instruments: the Question Appraisal System (QAS). The QAS is defined as a "taxonomy" of the cognitive demands of a question. It is a coding system based on four cognitive processes for response formation: comprehension, memory retrieval, judgement, and response selection (Tourangeau, Rips, & Kenneth, 2000). The results of the appraisal are used to revise question wording, questionnaire format and question ordering (Lessler & Forsyth, 1996). Although the system is useful for detecting the complexity of a survey item, it depends on the coders' ability to provide impartial judgements. The assessment includes many subjective categories such as if an item is difficult to read, if there are complicated instructions, or if a respondent is unlikely to

know an answer. If coders are used to technical language or are highly educated they could dismiss the complexity of a survey question.

Harkness and Schoua-Glusberg (1998) pointed out that assessment in survey translation is challenging because methods do not specify the criteria of assessment i.e. what is assessed and how. Saris (2012:548) reviews methods to evaluate survey questions and concludes that "all procedures based on personal judgments provide information about the validity, social desirability, and knowledge of the respondents about the issue of the question and much less about the effects of the form of the questions."

In other words, current procedures can be improved if the criteria for auditing a translation is specified and systematically monitored. Procedures based on judgements are not sufficient because evaluators look at different elements that matter for comparability, focusing on content but paying less attention to the effects of question wording and layout of questionnaires on equivalence. Pilot studies or split ballot experiments are a pretesting strategy that have a direct link to measurement equivalence, but they are not affordable for most surveys.

2.6 Formal characteristics of a survey item

Thanks to many years of research, we know which item characteristics are likely to affect a measurement instrument. Starting in 1951, Payne's book on the art of survey question

formulation already considered the consequences of different question formats and answer scales (Payne, 1951).

This tradition evolved and included experimental research to show how responses change between different formulations of a same concept (Schuman & Presser, 1981; Sudman, Bradburn, & Schwarz, 1996). Research has also defined the cognitive processes behind a survey response (Schwarz, 2007; Sudman et al., 1996; Tourangeau et al., 2000) and how different characteristics of a question, for instance, qualifiers in answer scales, affect this cognitive process (J. A. Krosnick & Fabrigar, 1997; Saris, 1988). Research has shown that item characteristics –such as layout, question form, response scale, labelling of response options, don't know option, length of the interview, among many others- may increase or decrease item bias and method effects (Alwin, 2007; Költringer, 1995; J. A. Krosnick & Fabrigar, 1997; Saris & Gallhofer, 2014; Tourangeau et al., 2000).

A related line of research, measurement quality, made it possible to estimate to what extent observed answers change when specific characteristics in a survey item also change and how serious this is in terms of measurement error (Alwin, 2007; Andrews, 1984; Költringer, 1995; Saris & Andrews, 1991, 2004; Scherpenzeel & Saris, 1997; Scherpenzeel, 1995).

When survey questions are designed and later on when they are translated, researchers take decisions that we have called the

characteristics of survey items. Saris and Gallhofer (2014:29) made an "inventory" of those decisions (over 60). They developed a coding scheme for this inventory to collect comprehensive information about the characteristics of a survey item and use them as predictors for measurement quality – defined as the variance of the observed variable explained by the variable of interest (Saris & Gallhofer, 2014). Translation procedures can use this coding scheme to monitor equivalence across languages. If the characteristics of source and target survey items are coded and compared using this scheme, differences in the codes mean that features that research has shown affect equivalence are different across language versions. This procedure provides a simple way to assess language versions before data collection.This coding scheme is incorporated in the Survey Quality Predictor survey software (SQP). The next section summarizes the current inventory of item features in SQP as a brief introduction to the software.

2.7 Survey characteristics in SQP

In their inventory, Saris and Gallhofer (2014) have included a comprehensive list of features that scholars in survey methodology have identified as the characteristics that affect a survey item. It is not the objective of this paper to go further into how these characteristics affect survey responses. Specialised literature in this regard is available (Alwin, 2007; Dillman, Smyth, & Melani, 2011; cf. Saris & Gallhofer, 2014) and the codebook (available at

www.sqp.upf.edu) provides an in-depth definition of each of the survey properties included in the program.

The classification of item characteristics can be divided into two levels of decisionsto be taken: 1) features that are inherent to the topic of interest and cannot be changed by the questionnaire designer and 2) the characteristics that are the product of decisions taken by the researcher when the item is formulated. Within those two groups, the list of survey characteristics in the SQP coding scheme can be categorized into ten subgroups shown in Table 1 below.

Group 1, the characteristics of the trait, includes four codes. The 'domain' is determined by the topic of the research. The 'concept' is an abstract aspect that the question measures about the topic, such as a feeling, a judgment, an evaluation, et cetera. The choice of domain and concept determines 'other associated characteristics' t-hat are coded in SQP, such as the presence of 'social desirability', the 'centrality' of the topic in the mind of the respondents, and the 'time specification' of the survey items.

The second group of codes specifies the formal characteristics of the request. The coder gives information on the 'basic choice': if it is a direct or an indirect request or if there is no request (in a battery). It is also coded if there is a 'WH word' and its 'type', if it measures quantity, extremity, intensity, place, time, etcetera. The request for an answer is classified as 'interrogative question',

‘imperative question or instruction’, ‘declarative statement’ or ‘none of three’ (subsequent batteries items).

Other properties in this group are if ‘gradation’ is used, if the request is ‘balanced’, if there is an ‘encouragement to answer’, if there is ‘emphasis on subjective opinion’, if the request contains ‘information about the opinion of other people’, if it demands an ‘absolute’ or a ‘comparative judgment’ and, whether batteries of questions use ‘stimulus or statements'.

The third group of codes in SQP are the measurement properties of the response scale. The program asks to code which is the ‘basic form of the response scale’, options are ‘categorical’ when the number of categories is between 3 and 12; ‘yes/no answer scales or a dichotomous choice’; ‘frequencies’, where amounts such as percentages, time, probabilities are requested; ‘magnitude estimation’, when size of numbers indicates the opinion; ‘line drawing’ and, ‘more steps procedures’.

Depending on its basic form, the program asks for other specific characteristics of the response scale (group 4). Codes in group 5 ask about the presence of ‘instructions for interviewers and/or respondents’. Group 6 includes ‘additional information’ about the topic or the scale, such as, ‘extra motivations, information or definitions’. Group 7 has codes about the characteristics of the ‘introduction’ (if any) and its specific features. Group 8 asks about the linguistic complexity of the item using as indicators the ‘number

of sentences', of 'subordinated clauses', of 'words', 'nouns', 'abstract nouns' and 'syllables' in the request for an answer, the answer scale and in the introduction (if present). Group 9 is about the 'method of data collection' and the language of the survey. Finally, group 10 is about the layout and content of showcards or visual aid (if used).

Table 1. Summary of characteristics inventoried by SQP

Group	Specific characteristic	
	Features that are inherent to the topic of interest and cannot be changed during questionnaire design	
Group 1	About the trait	Domain
		Concept
	Associated with the trait	Social desirability
Group 2		Centrality of the topic
		Time specification
	Features that are decisions taken during questionnaire design	
Group 3	Formulation of the request for an answer	Trait requested indirectly, direct or no request and presence of stimulus (battery)
		WH word and what type of WH word
		Type of the request (interrogative, imperative question-instruction, declarative or none (batteries).
		Gradation
		Balance of request or not
		Encouragement to answer
		Emphasis on subjective opinion
		Information about the opinion of other people
		Absolute or a comparative judgment
Group 4	Characteristics of the response scale	Categories; yes/no answer scale; frequencies; magnitude estimation; line production and, more steps procedures.
	If the selection is "categories":	Number of categories
	Characteristics of labels:	Full or partial labels
		Labels in long or short text
		Order of labels
		Correspondence between labels and numbers
		Theoretical range of scales (bipolar or unipolar)
		Range of scales used
		Fixed reference points
		Don't know option
Group 5	Instructions	Respondent instructions
		Interviewer instructions
Group 6	Additional information about the topic	Additional definitions, information or motivation

Table 1 Cont.

Group	Specific characteristic	
Group 7	Introduction	Introduction and if request is in the introduction
Group 8	Linguistic complexity	Number of sentences
		Number of subordinated clauses
		Number of words
		Number of nouns
		Number of abstract nouns
		Number of syllables
Group 9	Method of data collection	
	Language of the survey	
Group 10	Showcards or visual aid	Categories in horizontal or vertical layout
		Text is clearly connected to categories or if there is overlap
		Numbers or letters shown before answer categories
		Numbers in boxes
		Start of the response sentence shown on the showcard
		Question on the showcard
		Picture provided.

As an illustration, consider this item taken from the ESS Round 5 source questionnaire in English:

If a violent crime were to occur near to where you live and the police were called, how slowly or quickly do you think they would arrive at the scene? Choose your answer from this card, where 0 is extremely slowly and 10 is extremely quickly.

Extremely slowly											Extremely quickly
0	1	2	3	4	5	6	7	8	9	10	11

When coded into SQP, the ‘domain’ of this request is about ‘local institutions’ and the ‘concept’ specifies it is a ‘judgement‘. Other item characteristics that can be coded in SQP are that it is a ‘direct

request' in an 'interrogative' format using a 'WH word' and with a 'balanced' concept because it shows the two poles 'slowly/quickly'.Regarding the response scale, it can be said that it is 'categorical' the 'number of categories' is 11, it is 'partially labelled'; labels are 'short texts' and it has 'three fixed reference points' because the qualifier 'extremely' denotes an absolute ending point in the scale and there is a 'neutral' category (5).

It can also be said that this request has an 'instruction for the respondent': '*Choose your answer from this card…'*, a 'definition for the scale': '*…where 0 is extremely slowly and 10 is extremely quickly'* and a 'don't know option' which is not explicitly shown but only registered. The list of characteristics (approximately 60) allows having a very detailed map of the formulation of the item regardless the language.

When looking at the codes across the same item in different languages, it is obvious that characteristics such as 'Domain' and 'concept' should be kept the same. If they are different, the questions are referring to different topics, but when the questionnaire is translated and its layout arranged, there are other characteristics reflected in SQP codes that national teams could vary. This variation will affect the equivalence with the source and other target versions. Therefore one can use these codes to detect differences in the formulation of a question in different languages.

A very simple illustration of how the SQP coding scheme would help to detect deviations across languages is the information that the 'linguistic characteristics' provide for comparing translated items. It is true that the number of words, syllables and subordinated clauses vary depending on the structure of each language. However, outliers can be detected using very simple thresholds, for instance, one can check the number of languages in which items are above (or below) one and two standard deviations from the mean number of words, nouns, syllables; or simply those which exceed the number of sentences. Without knowing the meaning, this indicates an additional complexity (or simplification) of the items that could easily be confirmed in terms of content with the translation teams.

2.8 A five-step procedure for comparing item characteristics across languages

For survey questions in different languages, one can check if their characteristics are the same when the questions are coded into a same coding scheme and the codes are compared. This makes it possible to compare the characteristics independent of the languages. It would be an oversimplification of translation procedures to suggest that they can be solely evaluated by comparing item characteristics'codes. However, the five-step procedure helps the teams involved in the translation and assessment steps to define a framework to evaluate a survey item combining translation and functional equivalencerequirements. As Harkness, Villar, et al. (2010) suggested, translation procedures should define the unit of translation (the survey item), define the

elements that should match (intended meaning and intended measurement properties) and solve problems from this perspective rather than centring the discussion on the level of words.

a) Introducing questions in SQP

The first step is to upload questions from the source and target languages into SQP software. This can be done by any user at no cost after signing up and logging into the program at sqp.upf.edu webpage. When coding, the program displays a help option on each screen indicated by a yellow box, which defines each item characteristic asked and gives examples (a complete codebook is also available in a PDF version).

b) Coding the source questionnaire

The information regarding item characteristics of the source questionnaire must be accurate because target versions will be compared against it. It should be coded independently by two individuals with deep knowledge about questionnaire design; differences should be reconciled in collaboration with a third individual who plays the role of a reviewer.

c) Coding a target questionnaire

The translated questionnaire should be coded by a proficient speaker of the target language, preferably an individual involved in the translation process.

d) Comparison of measurement properties

The codes of the characteristics of source items should be compared with those in the target language. Any differences should be clarified with coders first, to rule out coding errors in the target questionnaire. True differences in the codes should be reported to the translation team.

e) Interpretation of deviations and actions taken in the target text

The translation team should clarify any difference in the codes in terms of the definition of the features. In other words, it should justify the reasons behind a deviation in the item characteristics. Depending on the type of difference, they may fall into one of three categories as shown in Table 2. Each category results in a suggested action for the translated text.

Table 2. Categories for differences in the SQP codes for two languages

Type of deviations found (source vs. translation)	Action taken
A) A difference that cannot be warranted, for instance, a different number of response categories, leaving out a "don't know" option or/and an instruction for the respondent.	The translation should be amended
B) A difference that may or may not be warranted e.g. use of complete sentences in the scales instead of short texts. In some languages it is necessary, in others this may be a fact of stylistic choice	Amendments in the translation are recommended to keep the principle of functional equivalence in translation if the language structure allows keeping the item characteristic the same.
C) A difference in the linguistic characteristics or grammatical structure that is unavoidable	Amendments in the translation are recommended to keep the principle of functional equivalence if the language structure allows a formulation closer to the source questionnaire.

2.9 Questions evaluated in the ESS

The five-step procedure to compare the codes the source questionnaire and translated language versions was applied in a sample of questions from Round 5 (R5), Round 6 (R6) and Round 7 (R7) of the ESS as a last step of quality control in the translation procedure. A total of 102 questions have been evaluated. 34 questions include the topics "Trust in criminal justice", "Attitudes towards immigration", "Personal and social well-being", "Democracy" and "Political efficacy" and 68 are repetitions of them with a variation in the measurement properties designed for experimental purposes[4].

[4]The formulation of the items in the main and supplementary questionnaires as designed in the English Source version is available at http://europeansocialsurvey.org

Twenty-nine language versions were coded in SQP in at least one round. Participating languages were Albanian, Catalan, Croatian, Czech, Danish, Dutch, Estonian, Finnish, French, German, Greek, Hebrew, Hungarian, Icelandic, Lithuanian, Norwegian, Polish, Portuguese, Russian, Slovak, Slovene, Spanish, Swedish, and Ukrainian. Table 3 shows all the combinations of participating languages per country and round. In the third column, it reports whether differences in the codes between the source and the language version were found and the fourth column summarises if they were reconciled or not. The Table 3 shows that in the vast majority of cases, there was some differences in the target versions but that at the end of the process, differences were reconciled or partially reconciled in almost all cases.

Differences in the codes were first checked by the translation team to rule out mistakes in the coding. Differences were reported to the national coordinators in each country and they were asked about the reasons for the differences in the translation e.g. if it was a decision taken due to the characteristics of the language, if it was a cultural problem, if it was a mistake in the translation process, et cetera. To minimize deviations from the source questionnaire, recommendations were provided when changes to the translation were not fundamental to the structure of the language.

The subsections below summarise findings falling into the three categories of Table 2: a) A difference that cannot be warranted, b) a difference that may or may not be warranted, c) a difference in the

linguistic characteristics that is unavoidable. Examples in languages other than English are back translated to make them understandable for this chapter.

a) Category A: Differences that cannot be warranted

The first category of differences correspond to unwarranted deviations made when the questionnaire was formatted in the target language. Deviations in this group were prevented, the most common ones were differences in the layout of questions and show cards, omission of parts of the item and, differences in form of the response scales.

Table 3.Summary information of participating countries in SQP Coding

Country	Language	Differences in codes between the source and the language version			Differences reconciled		
		R5	R6	R7	R5	R6	R7
Austria	German	-	Yes	Yes	-	Yes	Yes
Belgium	Dutch	Yes	No	Yes	Yes	-	Yes
Croatia	Croatian	Yes	-	-	Yes	-	-
Cyprus	Greek	-	Yes	-	-	No	
Czech Rep.	Czech	Yes	Yes	-	No	Yes	-
Denmark	Danish	Yes	Yes	No	Yes	Yes	-
Estonia	Estonian	Yes	No	Yes	Partially	-	Yes
Finland	Finnish	Yes	No	Yes	Partially	-	Yes
France	French	Yes	Yes	-	No	Yes	-
Germany	German	Yes	No	No	Yes	-	-
Greece	Greek	Yes	-	-	Yes	-	-
Hungary	Hungarian	Yes	Yes	Yes	No	Partially	Yes
Iceland	Icelandic	-	Yes	-	-	Yes	-
Israel	Hebrew	Yes	Yes	Yes	Partially	Yes	Yes
Kosovo	Albanian	-	Yes	-	-	Yes	-
Lithuania	Lithuanian	Yes	Yes	No	Partially	Partially	-
Netherlands	Dutch	No	-	Yes	-	-	Yes
Norway	Norwegian	-	Yes	Yes		Partially	Yes
Poland	Polish	Yes	Yes	Yes	Partially	Partially	Partially
Portugal	Portuguese	Yes	-	-	Yes	-	-
Russia	Russian	Yes	Yes	-	Yes	Partially	-
Slovakia	Slovak	Yes	-	-	Yes		-
Slovenia	Slovenian	Yes	-	-	Partially		
Spain	Spanish	Yes	Yes	Yes	Partially	Yes	Yes
Spain	Catalan	Yes	Yes	-	Partially	Yes	
Sweden	Swedish	Yes	No	Yes	No	-	Yes
Switzerland	French	Yes	-	-	Partially		
Switzerland	German	-	Yes	Yes	-	Yes	Yes
Ukraine	Ukrainian	Yes	No	-	Yes	-	-

- Indicates that country did not participate in SQP Coding for that round

- Layout of questions and show cards

In four language versions, "stand-alone" questions were formulated as part of a battery (questions about a similar topic grouped into a set that use exactly the same response format) and in two, the

opposite happened: batteries were presented as items. A typical example of this difference is shown in Figure 5, it presents the items in the translated version and Figure 6 shows the layout as in the source questionnaire. Alwin (2007, Krosnick (1990), Neijens (1987), Sanchez (1992) and Saris and Gallhofer (2014) among others have found that batteries have an effect on the quality of responses to attitudinal questions. In batteries, the complexity between the first and the subsequent items is different. In addition, if the questionnaire was self-administered, it cannot be assumed that respondents understand the way in which the set of questions should be answered.

De Leeuw (2008) and Dillman (2007) argue that show cards reduce the cognitive burden for respondents, but there is little research on how the layout affects responses. Therefore, once the source design is decided it is important that show cards remain identical across language versions. Changes were prevented in a total of twelve language versions. They consist of the visual presentation of an answer scale in a vertical or horizontal format, overlap of numbers with labels of the scale, additional numbers in front of fully labelled answer scales, additional boxes to frame categories and additional show cards in items which were not supposed to have one.

- Layout of response scales

A common group of differences points out unintended mistakes in the way response scales are designed (seven language versions).

Literature on questionnaire design has shown exhaustively that subtle differences in labels, number of categories and in the non-response options have a large impact on responses (Paul P Biemer, Groves, Lyberg, Mathiowetz, & Sudman, 2011; Saris & Gallhofer, 2014). Among the differences in this group are the number of categories, additional spontaneous response options explicitly shown as part of the categories, additional 'No answer' option to the 'Don't know' and, additional labels for categories that were designed as with numbers only.

Finally, in some other cases, mistakes were made while formatting the questionnaire and unnecessary repetitions or incorrect stimuli was provided, making the items more complex (four cases).

b) Category B: differences that may or may not be warranted

In the process of comparing the characteristics of the English source questionnaire and the translated versions, some deviations could not be solved. For each deviation, there was feedback provided to the national team and received from it. The languages used in the questionnaires in a large scale cross-cultural survey may be very different from each other, their structures being closer or less similar to the source language. In some languages it is difficult to combine several properties which are quite common in English. This information has been quite useful for designing a better source questionnaire because problems rarely belong to one language but are common to families of languages.

- Missing parts in an item

A group of differences that can be warranted in some cases but not in others are about the exclusion of parts in the item (eighteen language versions). These differences are challenging because responses could change depending on the decision taken, but at the same time there are strong arguments to keep differences arguing not only grammatical problems but cultural or idiomatic differences

For instance in the question

'Based on what you have heard or your own experience, how unsuccessful or successful do you think the police are at preventing crimes in [country] where violence is used or threatened?'

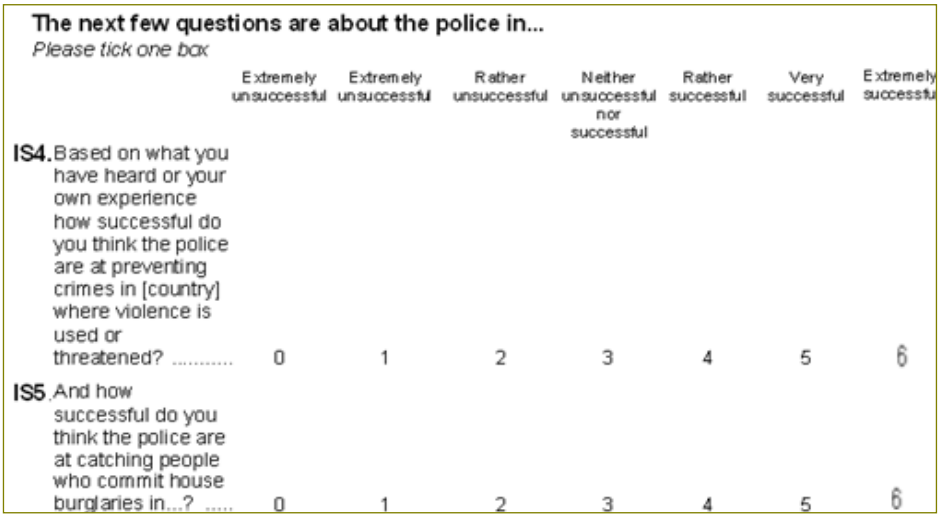
the definition of the scale

'Choose your answer from this card, where 0 means very unsuccessful and 4 means very successful.'

was left out, but it was asked to be included because the respondent could have interpreted the question as dichotomous. Other parts that were excluded were introductions to the questions, instructions for the respondents, omission of interrogative wording (WH words: '*how likely...?*', '*to what extent...?*'). If omitted, these elements introduce unintended differences in the measurement instruments

across languages, as the potential effect of the difference in responses is unknown, therefore in most cases they were asked to be incorporated.

Figure 5. Layout of coded items in thetranslated version



The next few questions are about the police in...

Please tick one box

	Extremely unsuccessful	Extremely unsuccessful	Rather unsuccessful	Neither unsuccessful nor successful	Rather successful	Very successful	Extremely successful
IS4. Based on what you have heard or your own experience how successful do you think the police are at preventing crimes in [country] where violence is used or threatened?	0	1	2	3	4	5	6
IS5. And how successful do you think the police are at catching people who commit house burglaries in...?	0	1	2	3	4	5	6

Figure 6. Layout of coded items in the source questionnaire

The next few questions are about the police in [country].

IS4 Based on what you have heard or your own experience how successful do you think the police are at preventing crimes in [country] where violence is used or threatened?[1] **Please tick one box.**

Extremely successful	Very unsuccessful	Rather unsuccessful	Neither unsuccessful nor successful	Rather successful	Very successful	Extremely successful
00	01	02	03	04	05	06
☐	☐	☐	☐	☐	☐	☐

IS5 And how successful do you think the police are at catching people who commit house burglaries in [country]? **Please tick one box.**[2]

Extremely successful	Very unsuccessful	Rather unsuccessful	Neither unsuccessful nor successful	Rather successful	Very successful	Extremely successful
00	01	02	03	04	05	06
☐	☐	☐	☐	☐	☐	☐

In another example, Hebrew, instructions for respondents such as '*use this card to answer'* have been left out. The national team has argued that in the natural course of the interview these instructions are not necessary, they make the interaction far longer and the interviewer can provide it in a more natural way.

- Polite forms in English

The source English language makes use of polite expressions that are cumbersome in some languages, among others, Swedish, Finnish, Norwegian or French. These expressions put emphasis on subjective opinion or encouragements to answer such as '*would you say*', '*please say what you think*', et cetera. Politeness in English can make an item imperative and indirect with the use of subordinated clauses, '*please tell me to what extent...*' rather than interrogative. In ten cases, language versions adapted the use of politeness to what is fluent in the target language. As an example, the question in the source questionnaire

Using this card, please tell me how interested you would generally say you are in what you are doing

was presented in a simplified version in Norwegian where the question is asked directly rather than in an imperative form

How interested are you generally in what you are doing? Use this card to answer.

The feedback given to the national team asked to look for a translation closer to the source version, if that would not compromise fluency in the language. In this case, the final version was formulated as:

Use this card to say how interested you are generally in what you are doing?

- Inconsistent translation in repeated questions, scale labels or instructions

Inconsistent translations in formulations such as the stem of the question, instructions or labels that were used in several items were very common. In thirteen cases, some variation was found in repeated questions. One example is the translation of *'violent crime'* which was formulated as *'aggression'* in a first occasion and as *'a crime or an offense'* in a repetition of the same concept.

In five other cases, labels in ending points of scales that should be the same, did not match. In five cases, repeated instructions for respondents had slight variations. Expressions that are used several times should be translated in the same way; there is no need to develop a new translation for instructions or parts of a question that are used several times in a questionnaire. This can be especially problematic in the design of experiments, where varying elements in the formulation of items disturbs an experimental design. A systematic check of repeated wording is difficult without a program. The coding system asks for number of words, nouns and abstract nouns in items. It is easy to detect a deviation if these

numbers are different in expressions that are repeated several times in a questionnaire. In this way, instructions or concepts can have a coherent translation in other parts of the questionnaire.

- Increased complexity of the items

Another frequent deviation was the addition of extra explanations, idiomatic expressions and instructions that made the items more complex but that national teams in some countries argued were necessary to add fluency or to make the question more understandable to respondents with low levels of formal education. An example of this is the instruction for the respondent

'Choose your answer from this card, where 0 is extremely unsuccessful and 10 is extremely successful'

was translated in Polish as

'*Please answer using this card and indicate the number from 0 to 10, where 0 is that it is completely ineffective to prevent such crimes, and 10 that it is fully effective. Other numbers are used to express an opinion in between'.*

Allowing this deviation in one country is problematic for a comparative survey, because if it is the case that the definitions of the scale are not sufficiently clear for all respondents (especially those with low levels of formal education) the same argument should apply for the rest of the

countries. If it is the case that additional explanations are needed in Poland, maybe they are needed everywhere and the source questionnaire should be modified.

- Formulation of labels in answer scales

In addition to the number of categories, the formulation of labels in answer scales has received attention in the questionnaire design literature (Alwin, 2007; J. A. Krosnick & Fabrigar, 1997; cf. Saris & Gallhofer, 2014). One of the main lessons learnt from comparing the characteristics of source and translated questions is regarding the huge challenges that national teams face in order to accurately translate scale labels. A particular challenge (seven cases) is in the use of short texts for labels. In some languages, it is grammatically incorrect (Balto-Slavic languages and Finnish are an example) to leave "alone" adverbs or adjectives (*completely-not at all*) without verbs or without grammatical persons (e.g. *extremely depressed, extremely lonely*). Thus, it is incorrect to use short texts for labels without more context such as verbs or pronouns, in these cases, translation teams have to use full sentences. However, short labels were not only problematic due to language structures. Some national teams argued that short texts are not clear enough and may cause confusion for respondents with low levels of formal education. As an example, in the question

'If a violent crime were to occur near to where you live and the police were called, how slowly or quickly do you think they would arrive at the

scene? Choose your answer from this card, where 0 is extremely slowly and 10 is extremely quickly'.

Extremely slowly										Extremely quickly	
0	1	2	3	4	5	6	7	8	9	10	11

the scale labels were formulated as *'It will arrive extremely late to the place"* and *"It will arrive extremely quickly to the place."* The main argument was that *'Extremely slowly/quickly'* could be interpreted as "driving fast or slow."

Another frequent deviation was the translation of qualifiers defined as *fixed reference points*. A fixed reference point is an anchor; there is no doubt about its position on the subjective scale in the mind of the respondent (Saris, 1988). The source English questionnaire in the ESS makes extensive use of the anchor 'Extremely' and this word was challenging for translation especially into Balto-Slavic languages. When this problem first appeared, national teams argued that there was no equivalent adverb to '*extremely'*, thus they reformulated the labels of the end-points as *'very*'. This second form is not a fixed reference point but a vague qualifier because respondents can have a different idea of what *'very'* means depending on their own subjective reference. A second argument given by the countries was that the formulation was possible but it was difficult to be understood by less educated people because it could be understood as *'extremist'.*

This flagged a recurrent problem in the translation of labels: national teams looked for a literal translation. This problem has decreased

substantively after a better communication with the translation teams regarding the intended measurement properties that the scale should convey (annotations were given that if extremely was not possible, the terms fully, absolutely, completely, totally have the same properties).

Another example of an incorrect translation of fixed reference points was found in four languages: Czech, Slovakia, Russia and Poland translated the scale *'not at all likely'* as *'very unlikely'* and the scale *'not at all often'* into *'very rare'*. The reason was that the expression *not at all* is idiomatic in English. Therefore, it is difficult to decide how it can be represented in target languages. A solution was to use the equivalent to 'never' instead of trying to reproduce '*not at all*'.

This process has helped to improve the guidelines and communication with countries on the elements that should be dealt with when the translations are done. Problems with fixed reference points were found in nine cases in Round 5, in six cases in Round 6 and only in one case in Round 7.

c) Category C: Differences in the linguisticcharacteristics

- Missing parts in an item

There are cases where the difference is inherent to the way the language is structured. In Lithuanian, for example, it is not appropriate to use a question word *'how'* and at the same time to use the two poles of the scale, '*slowly or quickly'*. The expression *'kaiplėtaiargreitai (*how slowly

or quickly) is not appropriate in the language. In order to keep a request balanced, the translation decision was to omit the WH word (*doesit arrive slowly or quickly*). They could include the WH and omit either *"quickly" or "slowly"* resulting in an unbalanced request.

- Bipolar vs. unipolar scales

A second common challenge appears in the choice of the appropriate translations to express antonyms when, in the source English questionnaire, they are expressed with the aid of the prefix 'un' (unimportant-important, unsuccessful-successful, able-unable, and confident-unconfident). Several translation teams interpreted these adjectives as bipolar. For instance, in the case of *'unsuccessful-successful*', they argued that as there was not an equivalent formulation in a bipolar range, they must adapt the concept. These qualifiers were translated as *'inefficient/efficient', 'ineffective/effective'* and *'bad/well'* in Spanish and Catalan, French and Finnish respectively. In other cases, translations were formulated ranging from *not successful* to *extremely successful* being a unipolar scale.

2.10 General discussion

This chapter focuses on a current problem in comparative survey research. Survey translation has developed best practice procedures to translate functionally equivalent survey questionnaires but, in practice, it is a very complex challenge to empirically check that the requirements set by translation guidelines are fulfilled. The elements that matter to design a

good question should be monitored in all participating languages. This is unworkable without a program and a clear inventory of the elements that should remain fixed across languages.

We reviewed best practice procedures to translate and assess translations of survey items arguing that they do not have a direct link to testing equivalence. They rely on judgements that may be partial, or focusing only on some characteristics or cognitive processes, or subjective, because they depend on the evaluators' knowledge of the context of the survey or even stylistic preferences about the language. Procedures that are thought to test equivalence, such as pilot studies, or split ballot experiments are not affordable for most survey projects or are only possible for a limited number of languages participating in a large scale comparative survey.

We suggest that future research should be addressed along two lines: the first to strengthen the link between functional equivalence and translation assessment. Translation assessment procedures need to be developed keeping in mind a framework for statistical equivalence. The second line of research is about how to manage large amounts of information and problems derived from translation procedures. Large scale cross-cultural survey projects involve many different languages and translation problems can be generalized into families of languages or they can be very specific for a language. This makes it very difficult to systematize information in survey translation without the aid of software. Survey translation may find it useful to explore solutions that are emerging in

other areas of translation such as the use of corpora or computational linguistics.

We suggested a procedure to detect deviations that are relevant for comparability of different language versions of a survey instrument before it is administered to respondents. It requires comparing the item characteristics of source and target survey items in a same coding scheme. This coding scheme is integrated in the form of a semiautomatic software called Survey Quality Predictor (SQP). Once survey items are coded into SQP, their characteristics can be compared in a systematic way.

By defining the intended measurement properties of a survey item that should remain constant across countries, the process was successful at preventing a large number of differences that were not warranted and has helped to better communicate the objectives of survey translation. This led not only to changes in the translation of some items in some languages, but also to better annotations of the source questionnaire and to fewer idiomatic expressions. The result is that, in general, the form of the target questions is closer to the source questionnaire. The process has a clear strength in that it structures the communication and feedback between questionnaire designers and translation teams. As questionnaire designers, we are now better aware of the challenges that national teams face to follow the ASQ model.

There is a group of differences between the source and target versions which is unavoidable due to the structure of target languages. For instance, some languages cannot leave alone short texts in the scales such

as adverbs without a complement or a grammatical person. The process has helped to be aware of those differences and to aid translators in taking better decisions when the form of the question should be adapted.

Finally, there were some differences which, in principle, could be avoided, but they were not reconciled because translation teams had strong cultural arguments to keep them. For instance, that additional explanations and definitions are helpful to less educated people and that fluency needs to be improved in the interview or that some qualifiers have a negative connotation in the language ('*extremely*'). There is not a final answer to this issue. It opens a line of debate for further research on comparative questionnaire design.

SQP Coding is not exempt from limitations. The first is inherent to coding procedures: coding can be tiresome and coders should be trained carefully to minimize coding errors. The second is about the inventory of characteristics that are compared, throughout the application of the procedure in three rounds of the ESS. We have learnt that some categories could be added. The third limitation is in the scope of the approach, as the codes focus on the form of the items content is dismissed. This issue has been tackled by asking national teams during the reporting step how specific formulations were solved, e.g. the dichotomy unsuccessful-successful, unimportant-important.

Chapter 3

Cross Cultural or Cross National Research? The Role of Language in a Comparative Survey

3. Cross Cultural Or Cross National Research? The Role Of Language In A Comparative Survey[5]

Abstract

This paper examines whether or not linguistic groups exhibit invariance within countries. It tests configural, metric and scalar invariance across linguistic groups in a model that distinguishes the response process - taking into account systematic error components- and the cognitive process. Our findings show that when differences in the response process are allowed, concepts are (partially) invariant across groups. The analysis is conducted for items measuring trust in institutions and political satisfaction for six linguistic groups including French, Dutch, Estonian, German, Ukrainian and Russian in four countries. Data comes from the European Social Survey.

3.1 Introduction

Previous research has established that measurement invariance (MI) is a prerequisite for deriving substantive conclusions for comparisons of data collected in diverse populations (cf. Davidov et al., 2014; Horn & McArdle, 1992; Meredith, 1993; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). (Horn & McArdle, 1992, p. 117) defined MI as "whether (or not) under different conditions of observing and studying phenomena measurement operations yield measures of the same attribute". A well-established procedure to test for MI is an analysis of the

[5] Submitted for publication and under review at the time of the dissertation's defense.

mean and covariance structure of latent variables. It is done using confirmatory factor analysis (CFA) within the framework of multi-group structural equation modelling (MG-SEM).

In addition to the challenges of estimation and model testing (Cheung & Rensvold, 2002; Saris, Satorra, & Van der Veld, 2009; cf. Vandenberg & Lance, 2000), the standard approach for testing MI in comparative survey research faces a conceptual problem related to the definition of the groups that are compared (Byrne & Watkins, 2003). This is due to the fact that in most cases countries are the ultimate unit of interest, measurement invariance is tested at the country level assuming that they are homogeneous cultural entities. In multilingual countries, it is plausible to hypothesize that invariance can be rejected for different linguistic groups.

This paper focuses on the MI to survey questions taking into account linguistic diversity within countries. The first objective is to challenge a current practice to test for invariance at the country level without testing whether linguistic groups are invariant or not in countries where survey instruments are translated in more than one language. A second objective is to test for invariance distinguishing the response and cognitive processes to a survey question. It is argued that there could be differences in respondents' reactions to the formulation of the measurement instrument across groups. If they are controlled, comparisons across groups can be done when there is invariance in the parameters representing the way respondents interpret survey items.

The overall objective is to provide evidence which supports two general rules: first, to test for invariance across linguistic groups before aggregating data at the country level when there is more than one language in the questionnaire and second, to account for measurement error in the test. Empirical results in this paper show that by modifying the classical MI model to account for measurement error invariance can be achieved.

The structure of this paper is as follows: in Section 3.2, we summarize the role of language in a survey interview. In Section 3.3, we explain how invariance is typically tested in multilingual countries and we show an alternative model which distinguishes the response and cognitive components of the measurement process. Section 3.4 presents the data, model fitting and model testing procedures for the models specified in Section 3.3. In Section 3.5, the results of the empirical analysis are shown and in Section 3.6, we conclude and discuss the results.

3.2 The effect of language in a survey

Large scale survey projects across the globe e.g. the European Social Survey (ESS), the Programme for International Student Assessment (PISA), the European Values Survey (EVS), among others, translate their instruments into more than one language when a country has minority languages (European Social Survey, 2014, p. 6; European Values Survey, 2010, p. 121; PISA, 2010, p. 5).

Translating questionnaires into minority languages has pros and cons: the advantage is that non-response bias is prevented, in a country with two (or more) linguistic groups, it may be the case that one cultural group is not proficient into the other's language e.g. Russian speakers in Estonia do not necessarily speak Estonian. However, the translation of questionnaires into minority languages also faces challenges, for instance, the language of the questionnaire threatens comparability because survey items can be biased by translation decisions (Janet A. Harkness, Villar, et al., 2010; Van de Vijver & Leung, 1997; Van de Vijver & Tanzer, 2004). However, accurate translations do not warrant MI.

Research in cross-cultural psychology provides evidence to support a modern formulation of the Whorfian hypothesis in a survey interview: that the way linguistic forms operate in the mind of the respondent have an effect on how concepts are related (Hunt & Agnoli, 1991; Richard & Toffoli, 2009). In survey research, the Whorfian hypothesis can be formulated by stating that the language in the interview may affect the response process to survey items and/or the interpretation of questions.

A *survey item* can be defined as a request for an answer about a concept and a topic asked using a *measurement method* i.e. the combination of characteristics that define the formulation and administration of the request, such as the response scale, the mode of data collection, the use of showcards or visual aid, the translation procedure, the selection and assignment of languages, the introduction, the additional explanations, etcetera (Saris & Gallhofer, 2014). The *measurement process* of a survey item is a combination of the *cognitive process* - the process that makes the

respondent understand and interpret the concept that is asked in the form of an item- and the *response process* - how the respondent responds to the way the item is formulated, the reaction to the measurement method (Van der Veld, 2006).

When measurement invariance is tested using data aggregated at the country level a strong assumption is made: that the responses obtained from asking two questions in different languages administered to a multicultural sample are equivalent. This, in general, would be the case if the understanding/interpretation of the questions were the same and the measurement method affected the response process in the same way. However, it is not necessarily the case that two linguistic groups define concepts in a same way only because they are part of the same country.

Language is part of what defines our cultural identity (Cohen, 2009) and cultural practices (Schwartz, Unger, Zamboanga, & Szapocznik, 2010). It is related to how we assign meaning and retrieve information in order to answer questions (Peytcheva, 2008). Language has an effect on the measurement process of survey questions by activating the associations of the cultural frames that yield a response (Bond & Yang, 1982; Luna et al., 2008; Yang & Bond, 1980). This means that there can be different interpretations of questions depending on the cultural frames activated by the language in the survey, but it may also imply that respondents could react differently to the measurement method.

Language effects in the response and cognitive process can be identified by analyzing the response patterns of bilingual individuals. They may

switch their cultural frameworks depending on whether or not they are exposed to stimuli related to one language or the other, their reactions becoming closer to the cultural features of the stimuli's language. This phenomenon has been called Cultural Frame Switching (CFS) (Hong et al., 2000). Benet-Martinez et al. (2002) showed that the stimuli posed by language activated CFS when respondents answered to an instrument to measure personality traits. Later evidence has shown that CFS depends on the potential level of conflicted characteristics and the distance of the cultural frameworks which define an individual's identity (Benet-Martínez & Haritatos, 2005).

Equivalence in a multilingual survey is at stake, not only if the comprehension of questions depends on the cultural frameworks but also when the characteristics of the survey method affect respondents' reactions regardless of their opinion on a topic. Harzing (2006) found that for agree/disagree (A/D) items[6] answering a questionnaire in one's native language was related to a higher extreme-response style (the tendency to select systematically the extreme categories of the scale). She also found that English language proficiency was related to a lower middle-response style (systematic choice of the middle category) (Harzing, 2006). Potential implications of these results in countries with high migration rates are that respondents would use different response processes: first-

[6] We prefer refer to "agree/disagree" items instead of Likert items, because Likert scaling is in itself a psychometric procedure to design a measurement instrument. It is different to what is typically referred as a Likert scale i.e. a design where respondents are presented a battery of statements; they should express their opinions by agreeing or disagreeing with each statement. The scale for these items is typically five-point categorical (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree). For more details on the Likert scaling procedure see: Likert, Rensis (1932). "A Technique for the Measurement of Attitudes". Archives of Psychology 140: 1–55.

generation migrant respondents may use systematic patterns of response style which may be different to those associated to second or later generations or native population.

Uskul, Oyserman and Schwarz (2010) showed that the response process is highly influenced by features of the questionnaire and the survey context (cultural characteristics in which the survey takes place). They suggested that the emphasis on honour and modesty, highly valuable in some Latin American, Middle East and Mediterranean cultures, are particularly influential cultural traits in the response process. Harzing (2006) analyzed 26 cultural groups (defined by language) and found that cultural characteristics at the group level such as those associated with the dichotomy collectivism/individualism, power distance, uncertainty avoidance and extraversion produced response bias such as acquiescence i.e. the tendency to agree with a statement regardless of the content and, extreme/middle response style. The implication in multilingual countries is that if linguistic groups differ in their cultural characteristics, it is plausible that they observed differences in the response process.

Several authors have tested for invariance/equivalence in measurement instruments across linguistic groups defined as a combination of language and country or language and ethnic origin. They have provided evidence that MI cannot be assumed for multicultural samples when respondents are given the option to select the language of the interview (Keysar, Hayakawa, & An, 2012; Richard & Toffoli, 2009; Schwartz et al., 2014; Zavala-Rojas, 2012). However, these studies are limited because they have not explored the distinction between the response and the cognitive

processes. Therefore, it is not yet clear if the measurement invariance was because linguistic groups had a different understanding of the concepts or because the reaction to the measurement method was different, or both.

Richard and Toffoli (2009) found that for items measuring ethnic identity and acculturation in a sample of Greek-Canadians, invariance was rejected when bilingual respondents were given the choice of the language of the interview. Individuals who spoke both languages translated in the survey and selected in which to answer likely incurred in self selection bias.

Zavala-Rojas (2012) tested invariance in items measuring trust in institutions for four rounds in the ESS across all participating countries including those with two or more linguistic groups (two or more questionnaires) such as Belgium, Ukraine, Estonia, Switzerland, Finland and Israel. Results show that if one linguistic group was not invariant and the deviations were very large with respect to the other linguistic group within a country, metric or scalar invariance was rejected when the data was aggregated at the country level.

Schwartz et al. (2014) tested invariance on a measurement instrument for acculturation and found full metric and partial scalar invariance only when the language of the questionnaire was randomised for bilingual individuals with approximately the same proficiency in Spanish and English, higher education rates and living in a bicultural context. They argued that aggregating data before confirming invariance may threaten comparability when respondents are able to choose the language of the

questionnaire because 1) respondents are likely to switch cultural frames to answer the interview, 2) they may select a language based on fear of a stereotype threat e.g. a pressure to choose a language related to the mainstream ethnic group or region where the survey takes place even if the respondent is not sufficiently proficient on it, 3) they may lack linguistic competency to choose a language (possibly derived from 2) or, 4) the survey design may lack translation quality.

In general, research which has investigated invariance within a country for different linguistic groups concluded that it was not possible to aggregate at the country level. This evidence challenges current practices in comparative survey research where invariance is tested at the country level in multilingual populations. But past research does not tell us if it is a different response process, which introduces variation in measurement error across groups, or if cultural differences lead to differences in the interpretation of the concepts asked causing a different cognitive process, or possibly both. In the next section, we formally define the measurement invariance test in its classical form and an alternative model proposed by Saris and Gallhofer (2014) which allows us to distinguish the cognitive and response processes for linguistic groups within a country.

3.3 Testing invariance in multilingual countries

Measurement invariance is formally tested by restricting the parameters in the measurement models across groups to be equal. Equation (11) to

Equation (14) show a typical model[7] for the relationship between three observed variables y_i^g,i=1, 2, 3, and a concept by postulation[8]; ξ_j^g, j=1, for group g, g=1, 2, …, p (Bollen, 1989; Meredith, 1993). In this model,y_i^g is a measure of a concept by intuition in the form of a survey item (observed variable). In these equations, t_i^g represents the intercept of the regression; l_{ij}^g represents the slope (or factor loading) between the i_{th} observed variable and the j_{th} concept-by-postulation and e_i^g represents the disturbance term of the equation. It is assumed that the disturbance terms have a mean of zero, they are uncorrelated with each other and with ξ_j^g,represented by (14).

$$y_1^g = t_1^g + l_{11}^g \xi_1^g + e_1^g \quad (11)$$

$$y_2^g = t_2^g + l_{21}^g \xi_1^g + e_2^g \quad (12)$$

$$y_3^g = t_3^g + l_{31}^g \xi_1^g + e_3^g \quad (13)$$

$$E(e_i) = 0; E\left(e_i \xi_1\right) = 0; E(e_i e_{i\prime}) = 0 \text{ for} i \neq i' \quad (14)$$

[7] We show the relationship between three observed variables and one latent variable, but the response function can be generalized to $y_i^g = t_i^g + l_{ij}^g \xi_j^g + e_i^g$ where i=1, 2, …, p; j=1, 2, …,q and g=1, 2, …,n.

[8] Saris and Gallhofer (2014:15) make the distinction between concepts-by-intuition and concepts-by-postulation following Blalock (1990) and Northrop (1947). In this distinction concepts by intuition are "more or less immediately perceived by our sensory organs (or their extensions) without recourse to a deductively formulated theory". Concepts by postulation are constructs which require explicit definitions and are defined by simple concepts already understood or concepts by intuition. They are complex concepts that cannot be measured with only one but with several concepts-by-intuition. Under this definition, how satisfied are you with the government? and how satisfied are you with democracy in your country? are questions for concepts by intuition, whereas 'satisfaction with politics' is a concept-by-postulation made up by the former two.

Invariance in the parameters of the measurement model in Equation (11) to Equation (13) is tested in three steps, where each step is a prerequisite of the next one. In the first step, a *configural* model is fitted for all groups to check if the configuration of the factorial structure, the pattern of fixed and free parameters- is the same across the groups of interest. In the second step, *metric invariance*, the configural model is restricted to one where the factor loadings are invariant across the groups ($l_{ij}^{1} = l_{ij}^{2} = \cdots = l_{ij}^{p}$). When the model is not rejected, comparisons of relationships across groups can be made (Horn & McArdle, 1992). The third step, *scalar invariance* implies that in addition to invariance in the factor loadings, intercepts are also restricted to be the same ($t_i^1 = t_i^2 = \cdots = t_i^p$). If the model is not rejected, comparisons of means can also be made across groups. This standard parameterization is represented in Figure 7.

The standard procedure to test for invariance is too restrictive in the sense that it does not allow separation of the respondent's reaction to the measurement method, the *response process*, from a true difference in the interpretation of the meaning of concepts, the *cognitive process*. This limitation of the standard procedure has been referred to as *susceptibility* i.e. to what extent the procedure is sensitive to artefacts in the response process (Butts et al., 2006; Byrne & Watkins, 2003; Marsh & Byrne, 1993; Saris & Gallhofer, 2014).

Saris and Gallhofer (2014:285) suggested that a test for invariance should *allow* correction for differences in the *response process* (correction for measurement error). MI could be tested using a model that makes the

distinction between the *cognitive* and the *response* components, where only the parameters of the cognitive equations should be necessarily constrained to be equal across groups. If latent variables are used, comparisons across groups can be done when the meaning of the concepts is the same, even if the response process is not the same.

In the context of multilingual samples, as it is possible that each linguistic group has a different reaction to the method used in the questionnaire, one can allow differences in the response process and test only for differences in the interpretation of the concepts across groups. In Figure 8, the response process is modelled by the relationship between the concept-by-intuition, η_j^g, with each observed variable, y_i^g, where λ_{ij}^g is the slope of this relationship and τ_i^g, the intercept. At the upper part of the figure, the cognitive process is represented by the relationship between η_j^g, the concepts-by-intuition and ξ_k^g, the concept-by-postulation with γ_{ik}^g representing the slope of this relationship and α_j^g, the intercept.

As formulated in the figure, this model is not identified. To estimate a model distinguishing the cognitive and the response component, it is necessary to add one observed variable to measure each concept by intuition. This is shown in Figure 9. Finally, as the concepts-by-intuition are measured for the same respondents at two points in time and the only difference is the method used, two method factors should be added (η_4^g,η_5^g) representing the respondents' reaction to the measurement method as shown in Figure 10. The model in Figure 10 is the baseline model to test for MI across linguistic groups. This final model is

represented by Equation (15) to Equation (25). Standard constraints (26)-(29) were imposed to identify the baseline model. First, the loadings of method factors $\lambda_{i4,i5}^{g}$, were fixed to 1. Second, one of the loadings in the measurement equations was fixed to 1 to fix the latent scales (λ_{11}^{g}, λ_{22}^{g}, λ_{33}^{g}). The intercepts, τ_{i}^{g}, were all fixed to zero for the response equations. The first intercept of the cognitive equations, α_{1}^{g}, was also fixed to zero.Equations (24) to (25) represent standard assumptions, that the error terms, unique components and the method factors have a mean of zero and that they are not associated with each other or with the latent variables η_{j}^{g} and ξ_{k}^{g} in the model.This model will be used for estimation in the empirical section of this paper.

Response process

$$y_{1}^{g} = \tau_{1}^{g} + \lambda_{11}^{g}\eta_{1}^{g} + \lambda_{14}^{g}\eta_{4}^{g} + \epsilon_{1}^{g} \quad (15)$$

$$y_{2}^{g} = \tau_{2}^{g} + \lambda_{22}^{g}\eta_{2}^{g} + \lambda_{24}^{g}\eta_{4}^{g} + \epsilon_{2}^{g} \quad (16)$$

$$y_{3}^{g} = \tau_{3}^{g} + \lambda_{33}^{g}\eta_{3}^{g} + \lambda_{34}^{g}\eta_{4}^{g} + \epsilon_{3}^{g} \quad (17)$$

$$y_{4}^{g} = \tau_{4}^{g} + \lambda_{41}^{g}\eta_{1}^{g} + \lambda_{45}^{g}\eta_{5}^{g} + \epsilon_{4}^{g} \quad (18)$$

$$y_{5}^{g} = \tau_{5}^{g} + \lambda_{52}^{g}\eta_{2}^{g} + \lambda_{55}^{g}\eta_{5}^{g} + \epsilon_{5}^{g} \quad (19)$$

$$y_{6}^{g} = \tau_{6}^{g} + \lambda_{63}^{g}\eta_{3}^{g} + \lambda_{65}^{g}\eta_{5}^{g} + \epsilon_{6}^{g} \quad (20)$$

Cognitive process

$$\eta_{1}^{g} = \alpha_{1}^{g} + \gamma_{11}^{g}\xi_{1}^{g} + \zeta_{1}^{g} \quad (21)$$

$$\eta_{2}^{g} = \alpha_{2}^{g} + \gamma_{21}^{g}\xi_{1}^{g} + \zeta_{2}^{g} \quad (22)$$

$$\eta_{3}^{g} = \alpha_{3}^{g} + \gamma_{31}^{g}\xi_{1}^{g} + \zeta_{3}^{g} \quad (23)$$

Assumptions

$$E(\epsilon_i) = 0;\ E(\epsilon_i \eta_j) = 0;\ E(\epsilon_i \xi_k) = 0;$$

$$E(\epsilon_i\, \zeta_j) = 0;\ E(\epsilon_i \epsilon_j) = 0 \text{ for } i \neq j \quad (24)$$

$$E(\zeta_j) = 0;\ E(\zeta_j\, \xi_k) = 0;$$

$$E(\zeta_i \zeta_j) = 0 \text{ for } i \neq j;\ E(\eta_i \eta_j) = 0 \text{ for } i=1,2,3\ ,\ j=4,5 \quad (25)$$

Identification/initial restrictions

$$\lambda_{11}^{g} = \lambda_{22}^{g} = \lambda_{33}^{g} = 1 \qquad (26)$$

$$\lambda_{i4}^{g} = \lambda_{i5}^{g} = 1 \qquad (27)$$

$$\tau_1^{g} =, \dots, = \tau_6^{g} = 0 \qquad (28)$$

$$\alpha_4^{g} = \alpha_5^{g} = 0 \qquad (29)$$

Figure 7- Figure 10. Standard vs. alternative models to test for measurement invariance

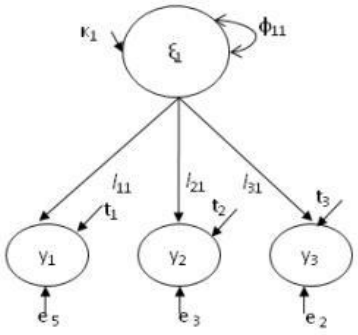


Figure 7

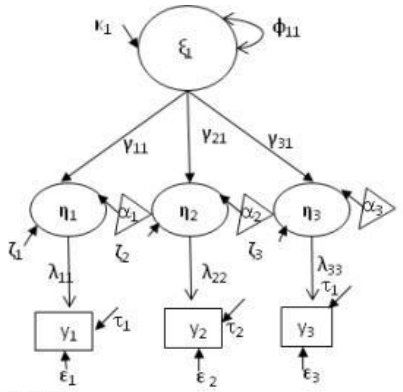


Figure 8

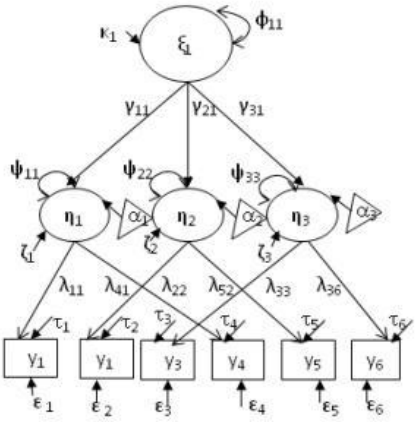


Figure 9

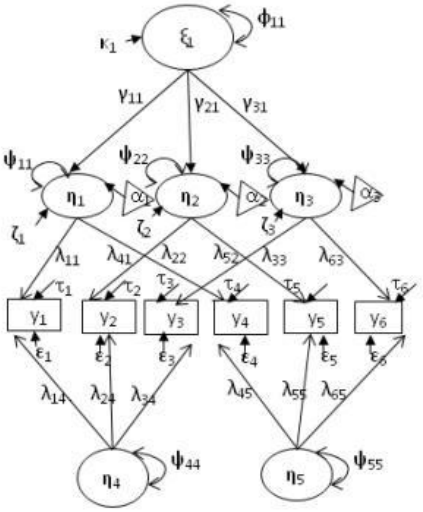


Figure 10

In the standard approach, the relationship between ξ_k^g and y_i^g is direct. In the alternative parameterization it is mediated by an intermediate response step represented by the bottom part of Figure 10 (Equation 15 to Equation 20). Variations in the parameters $\lambda_{i1}^g, \lambda_{i2}^g$ and or λ_{i3}^g are plausible, because respondents across cultural

groups have different reactions to the measurement procedures, regardless of their opinion about the concepts.

The top part of Figure 10 corresponds to the cognitive process which establishes the relationship between the concepts-by-intuition η_j^g and the concept-by-postulation represented by ξ_1^g, where ζ_j^g is a unique component of each concept-by-intuition with variance ψ_j^g. These relationships are assumed to be linear as shown by Equations (21) to (23). The loadings ($\gamma_{11}^g, \gamma_{21}^g, \gamma_{31}^g$) represent the slope of the relationship between $\xi_1^g$and η_j^g. The intercepts of the cognitive equations are represented by α_j^g.

The restrictions across the groups to test the null hypothesis of invariance are made in the cognitive equations (21)-(23): firstly by restricting the baseline model to be the same across groups (configural invariance); secondly by restricting the loadings of the cognitive equations to be the same to test for metric invariance ($\gamma_{41}^1 = \gamma_{41}^2;\ \gamma_{52}^1 = \gamma_{52}^2;\ \gamma_{63}^1 = \gamma_{63}^2$) and finally, restricting the intercepts ($\alpha_2^1 = \alpha_2^2;\ \alpha_3^1 = \alpha_3^2$) to test for scalar invariance. These restrictions test if cultural groups understand/interpret concepts in the same way, at the same time that the differences in the response process are taken into account. The response equations (15)-(20) may be estimated without constraining the parameters, because the main interest is to establish whether or not the underlying understanding of the concepts is the same. If the parameters of the response process are also restricted, the model is testing in addition

whether the reaction to the method is the same across groups. The decision procedure to determine whether or not a fixed parameter was misspecified was done following the Judgment Rule approach by (Saris et al., 2009). If a misspecification was found significant, the loading/intercept was freed. It is described in detail in the subsection Model testing[9].

We argue that invariance is often rejected in the standard approach because the coefficients represent mixed parameters of a response component and a cognitive component. The alternative approach to test for invariance is less restrictive to detect relevant differences across (linguistic) groups. True differences in the meaning of concepts are being tested and, at the same time, differences in the way respondents react to survey questions are allowed. It is no longer assumed, but can be tested explicitly, that both the interpretations of a concept and the reaction to the survey method are or have to be equal. This is illustrated in Equation (20)-Equation (25). Substituting Equation (21) - (23) for the first observed variable into Equation (5) - (10) as shown in Equation (30) - (32) results in the parameterization of the standard procedure.

$$y_1^g = \tau_1^g + \lambda_{11}^g\left(\alpha_1^g + \gamma_{11}^g \xi_1^g + \zeta_1^g\right) + \lambda_{14}^g \eta_4^g + \epsilon_1^g \qquad (30)$$

[9] We fit the models using LISREL 8.7 (Jöreskog & Sörbom, 2004). A repository with the inputs and outputs from Lisrel 8.7 and the data used in the analyses in this paper is found at this link:
https://github.com/dianazr/CrossNationalCrossCultural

By rearranging terms, this equation can be simplified to Equation (32):

$$y_1^g = \tau_1^g + \lambda_{11}^g \alpha_1^g + \lambda_{11}^g \gamma_{11}^g \xi_1^g + \lambda_{11}^g \zeta_1^g + \lambda_{14}^g \eta_4^g + \epsilon_1^g \quad (31)$$

$$y_1^g = \quad t_1^g \quad + \quad l_{11}^g \ \xi_1^g + \quad e_1^g \quad (32)$$

Where,

$$t_1^g = \tau_1^g + \lambda_{11}^g \alpha_1^g \quad (33)$$

$$l_{11}^g = \lambda_{11}^g \gamma_{11}^g \quad (34)$$

$$e_1^g = \lambda_{11}^g \zeta_1^g + \lambda_{14}^g \eta_4^g + \epsilon_1^g \quad (35)$$

It is clearly seen that the parameters of the standard approach are complex, they are made up by different components that belong to the cognitive process and the reaction to the way the question is formulated.

When the alternative model is used, if only differences in the response process are found, then comparisons of means and relationships across groups could still be done by correcting for measurement error using either a latent variable approach (or in the composite scores). However, if differences in the cognitive process are found, they represent true deviations across groups on how a concept is defined and interpreted.

In the following section, we provide detailed explanations of the data and model testing procedures we used to test for invariance using the alternative parameterization summarised by Equation (15) to Equation (29). Thereafter, the model results are presented.

3.4 Data, model testing, and model results

a) Data

- Linguistic groups in the analysis.

The model represented by Equations (15)-(29) is fitted for two linguistic groups in Belgium, Estonia, Switzerland and Ukraine using data from the European Social Survey Round 2 (European Social Survey, 2005). Countries were selected because the proportion of respondents in minority linguistic groups is at least 25% covering a diverse range of languages: French, Dutch, German, Estonian, Russian and Ukrainian. A model was fitted for each concept-by-postulation ' trust in institutions' and 'political satisfaction' with the objective to compare linguistic groups within a country. In this way, the empirical part shows results of four ‘Studies’. Table 4 shows the number of respondents in each linguistic group who answered the same concept using a different measurement method twice.

Table 4. Number of cases included in each study

Country	Languages	Number of cases	Study label
Belgium	Dutch	N = 653	Trust in institutions
		N = 334	Political satisfaction
	French	N = 476	Trust in institutions
		N = 237	Political satisfaction
Switzerland	German	N = 1878	Trust in institutions
		N = 490	Political satisfaction
	French	N = 978	Trust in institutions
		N = 156	Political satisfaction
Estonia	Estonian	N = 954	Trust in institutions
		N = 431	Political satisfaction
	Russian	N = 324	Trust in institutions
		N = 137	Political satisfaction
Ukraine	Ukrainian	N = 635	Trust in institutions
		N = 217	Political satisfaction
	Russian	N = 705	Trust in institutions
		N = 260	Political satisfaction

- Survey items

Table 5 shows the formulation of the survey items in the source English questionnaire for trust in institutions (PARLIAMENT, LEGAL, POLITICIANS). Those three items were repeated in a supplementary questionnaire to a random subsample of respondents. The only difference in the measurement method at time 2 was that the items did not have a 'Don't know' option. The table also shows the formulation of the survey items about political satisfaction (ECONOMY, GOVERNMENT, DEMOCRACY). In time 2 the difference in the measurement method was a verbal label to the mid-category.

Table 5. Survey items measuring 'trust in institutions' and 'political satisfaction' in the ESS Round 2

Concept by postulation: Trust in institutions	
Variable name	Item formulation
PARLIAMENTa	Using this card, please tell me on a score of 0-10 how much you personally trust each of the institutions I read out. 0 means you do not trust an institution at all, and 10 means you have complete trust. Firstly… …the [country]'s parliament?
LEGALa	…the legal system?
POLITICIANSa	…the politicians?
Concept by postulation: Political satisfaction	
Variable name	Item formulation
ECONOMYb	On the whole how satisfied are you with the present state of the economy in [country]? Still use this card.
GOVERNMENTb	Now thinking about the [country] government, how satisfied are you with the way it is doing its job? Still use this card.
DEMOCRACYb	And on the whole, how satisfied are you with the way democracy works in [country]? Still use this card.
Scales	a) Measured in time 1 and time 2 on an 11 point scale with labels at the ending points 0= not trust at all, 10 = complete trust. The difference in the formulation is that in time one, there is the possibility of the spontaneous 'Don't know' option whereas in time 2 this possibility is not present. b) Measured in time 1 on an 11 point scale with labels at the ending points 0= extremely dissatisfied, 10 = extremely satisfied. The scale in time 2 one is an 11 point scale with labels 0= extremely dissatisfied, 5= neither dissatisfied nor satisfied and, 10 = extremely satisfied
Source: European Social Survey (2004) "Round 2 Source Questionnaire". London: Centre for Comparative Social Surveys, City University London.	

b) Model testing

Typically, when global fit indexes are used to reject a group as invariant, groups with the largest chi-square contribution are excluded until a model with an acceptable fit on several fit indices are in acceptable ranges (Reeskens & Hooghe, 2007; Steinmetz, 2011). Using this criterion Allum, Read, and Sturgis (2011) tested invariance of the measures of political and social trust in Rounds 1,

2 and 3 of the ESS where only twelve out of seventeen countries were accepted as invariant. They excluded countries with the largest Chi-square until they got acceptable fitted models with an RMSEA< .08 and SRMR<.05. Fit indices in structural equation modelling and of model rejection for invariance testing are quite controversial (F. F. Chen, 2007; Cheung & Rensvold, 2002; Saris et al., 2009), the Chi-square test and the RMSEA do not take into account Type II error (Saris et al., 2009); as a consequence, a misspecified model can be accepted whereas a model with irrelevant misspecifications can be rejected. Global fit measures are very strict because they only indicate acceptance or rejection of a model. They do not give an insight into which elements in a model are misspecified. The approach in this paper was to evaluate the models for local misspecifications instead of global fit indexes (we report global fit indexes in Appendix 3.1).

Saris et al. (2009) developed a procedure to determine whether misspecifications are present in SEM. The procedure tests directly for misspecifications in the model while taking into account the power of the test for each fixed parameter. A misspecification occurs if a parameter has been given a fixed value, which is incorrect in the population (Hu & Bentler, 1998). The misspecification test combines knowledge of: (a) the size of the misspecification (Expected Parameter Change, EPC); (b) the impact on the fit if the parameter was included as a free parameter (Modification Index, MI); and (c) the sensitivity of the test in detecting the misspecification (power of the test). Both (a) and (b)

are given by the outputs of SEM software; (c) can be calculated based on the non-centrality parameter. The program JRule (Van der Veld, Saris, & Satorra, 2008) facilitates the procedure by taking automatically (a) and (b) and estimating (c). In Table 6 the decision rules are presented based on this information.

Table 6. The decisions to be made in the different situations defined on the basis of the size of the modification index (MI) and the power of the test.

	High power	**Low power**
Significant MI	See whether or not the size of the EPC is larger than the threshold. Parameter set free if it is	Misspecification present (Parameter is freed)
Non significant MI	No misspecification (Parameter is not freed)	Inconclusive (Parameter is not freed)

To use this approach, one has to specify in advance which power is required to detect a misspecification for specific values of the parameters. A power of 0.8 was chosen to detect standardized loading differences larger than 0.1 and intercept differences larger than 0.05 times the length of the scales in the items. The JRule program was used to identify if fixed or constrained parameters were misspecified with respect to configural, metric and scalar invariance models in Equation (5) - Equation (15) (see Cieciuch et al. 2015; van der Veld and Saris 2011 for applications of Jrule as a model testing method to test for mesurement invariance). Parameters indicated as misspecified were freed until there were no more significant misspecifications.

c) Results of four studies

This section describes the results of testing for measurement invariance in trust in institutions and satisfaction with politics in four countries with minority languages. Table 7 to Table 10 show the coefficients of both the cognitive and the response process. Standard errors are shown in parentheses. When the parameters are invariant, the estimates are shown in one group and the label INV is attached to the other group.

- Study 1. Belgium

Table 7 shows the results for Dutch and French linguistic groups in Belgium. For trust in institutions, full scalar invariance was established, both at the cognitive and at the response levels. In the model for satisfaction with politics, full scalar invariance is established at the cognitive level but in the response process, one parameter of the first method factor was misspecified and unconstrained (λ_{34}). One intercept at the response level was also misspecified and freed (τ_1).

- Study 2. Switzerland

In Table 8, the coefficients of the models are shown for German and French in Switzerland. For trust in institutions, the findings are different than those observed in Belgium. Full scalar invariance is established at the cognitive level, but at the response level,

parameters λ_{14} and λ_{34} of the first method factor and λ_{45} of the second method factor were misspecified and freed. For satisfaction with politics, full scalar invariance was established at the cognitive level but at the response level parameter λ_{34}in the first method factor was misspecified.

- Study 3. Estonia

Table 9 shows results for Estonia. For trust in institutions, Russian and Estonian linguistic groups are invariant at the response level but they are partially scalar invariant at the cognitive level. The loading of the variable 'politicians', γ_{31}, was misspecified and freed. Therefore, the corresponding intercept was also unconstrained. In the case of satisfaction with politics, as in Belgium and Switzerland, the parameters were invariant at the cognitive level. At the response level parameters of the first method factor λ_{14} and λ_{24}were misspecified and unconstrained. An intercept at Time 1, τ_5, was also misspecified and freed.

- Study 4. Ukraine

For Ukraine, results are shown in Table 10. The parameters of trust in institutions are fully scalar invariant both at the cognitive and at the response level. The results of the model for satisfaction with politics are less similar to the other three countries. The parameters at the response level were misspecified, in this case, not only misspecifications in the method factors were found but also in the

loadings of the relationship between the concepts-by-intuition and the observed variables. Partial scalar invariance was established for the parameters at the cognitive level. The loading for the variable 'Economy', γ_{11}, was misspecified and freed, therefore he corresponding intercept, α_1, was freed as well.

Table 7 Measurement invariance test results for Dutch and French in Belgium

				Study 1. Belgium			
Cognitive process			Dutch	French		Dutch	French
Scalarinvariance							
ξ_1	Parliament	α_1	INV	0.00	Economy	α_1 INV	0.00
	Legal	α_2	INV	0.58 (0.16)	Government	α_2 INV	-1.79 (0.35)
	Politicians	α_3	INV	-0.80 (0.16)	Democracy	α_3 INV	-0.39 (0.34)
Metricinvariance							
ξ_1	Parliament	γ_{11}	INV	1.00	Economy	γ_{11} INV	1.00
	Legal	γ_{21}	INV	0.91 (0.03)	Government	γ_{21} INV	1.21 (0.06)
	Politicians	γ_{31}	INV	1.08 (0.03)	Democracy	γ_{31} INV	1.10 (0.06)
Response process							
Question at time 1g							
η_1	Parliament	λ_{11}	1	1	Economy	λ_{11} 1	1
						τ_1	-0.70 (0.12)
η_2	Legal	λ_{22}	1	1	Government	λ_{22} 1	1
η_3	Politicians	λ_{33}	1	1	Democracy	λ_{33} 1	1
Question at time 2							
η_1	Parliament	λ_{41}	INV	1.02 (0.01)	Economy	λ_{41} INV	0.97 (0.01)
η_2	Legal	λ_{52}	INV	1.01 (0.01)	Government	λ_{52} INV	1.03 (0.01
η_3	Politicians	λ_{63}	INV	1.01 (0.01)	Democracy	λ_{63} INV	1.00 (0.01)
Method factor 1							
	η_4	λ_{14}	1	1	η_4	λ_{14} 1	1
		λ_{24}	1	1		λ_{24} 1	1
		λ_{34}	INV	0.05 (0.15)		λ_{34} 1	0.36 (0.15)
Method factor 2							
	η_5	λ_{45}	1	1	η_5	λ_{45} 1	1
		λ_{55}	1	1		λ_{55} 1	1
		λ_{65}	INV	0.75 (0.11)		λ_{65} 1	1
Latentmeans		K_1	4.90 (0.08)	4.39 (0.09)		K_1 5.42 (0.10)	5.34 (0.12)

Table 8 Measurement invariance test results for German and French in Switzerland

		Study 2. Switzerland					
Cognitive process		German	French			German	French
Scalar invariance							
ξ_1	Parliament	α_1 INV	0.00	Economy	α_1 INV		0.00
	Legal	α_2 INV	1.65 (0.12)	Government	α_2 INV		0.11 (0.29)
	Politicians	α_3 INV	-0.16 (0.12)	Democracy	α_3 INV		1.64 (0.31)
Metric invariance							
ξ_1	Parliament	γ_{11} INV	1.00	Economy	γ_{11} INV		1.00
	Legal	γ_{21} INV	0.82 (0.02)	Government	γ_{21} INV		0.97 (0.06)
	Politicians γ_{31}	INV	0.90 (0.02)	Democracy	γ_{31} INV		0.90 (0.06)
Response process							
Question at time 1							
η_1	Parliament	λ_{11} 1	1	Economy	λ_{11} 1		1
η_2	Legal	λ_{22} 1	1	Government	λ_{22} 1		1
η_3	Politicians	λ_{33} 1	1	Democracy	λ_{33} 1		1
Question at time 1							
η_1	Parliament	λ_{41} INV	1.00 (0.00)	Economy	λ_{41} INV		1.05 (0.01)
η_2	Legal	λ_{52} INV	0.99 (0.00)	Government	λ_{52} INV		1.05 (0.01)
η_3	Politicians	λ_{63} INV	0.99 (0.00)	Democracy	λ_{63} INV		1.00 (0.01)
Method factor 1							
η_4		λ_{14} 1	1.63 (0.45)	η_4	λ_{14} 1		1
		λ_{24} 1	1		λ_{24} 1		1
		λ_{34} 0.29 (0.08)	1.05 (0.22)		λ_{34} 0.08 (0.12)		1.20 (0.53)
Method factor 2							
η_5		λ_{45} 1.71 (0.37)	0.79 (0.09)	η_5	λ_{45} 1		1
		λ_{55} 1	1		λ_{55} 1		1
		λ_{65} 1	1		λ_{65} 1.21 (1.10)		INV
Latentmeans		K_1 5.48 (0.04)	5.40 (0.06)		K_1 5.28 (0.08)		5.11 (0.16)

Table 9 Measurement invariance test results for Estonian and Russian in Estonia

Cognitive process			Study 3. Estonia				
			Estonian	Russian			Estonian	Russian
Scalar invariance								
ξ_1	Parliament	α_1	INV	0.00	Economy	α_1	INV	0.00
	Legal	α_2	INV	1.32 (0.13)	Government	α_2	INV	-1.53 (0.28)
	Politicians	α_3	-0.28 (0.14)	-0.11 (0.17)	Democracy	α_3	INV	-0.72 (0.27)
Metricinvariance								
ξ_1	Parliament	γ_{11}	INV	1.00	Economy	γ_{11}	INV	1.00
	Legal	γ_{21}	INV	0.85 (0.03)	Government	γ_{21}	INV	1.19 (0.06)
	Politicians	γ_{31}	0.87 (0.03)	0.76 (0.04)	Democracy	γ_{31}	INV	1.15 (0.06)
Response process								
Question at time 1								
η_1	Parliament	λ_{11}	1	1	Economy	λ_{11}	1	1
η_2	Legal	λ_{22}	1	1	Government	λ_{22}	1	1
η_3	Politicians	λ_{33}	1	1	Democracy	λ_{33}	1	1
Question at time 2								
η_1	Parliament	λ_{41}	INV	0.97 (0.01)	Economy	λ_{41}	INV	1.00 (0.01)
						τ_5	0.00	0.36 (0.10)
η_2	Legal	λ_{52}	INV	0.98 (0.01)	Government	λ_{52}	INV	0.96 (0.01)
η_3	Politicians	λ_{63}	INV	0.96 (0.01)	Democracy	λ_{63}	INV	1.00 (0.01)
Method factor 1								
η_4		λ_{14}	1	1	η_4	λ_{14}	1	1
		λ_{24}	1	1		λ_{24}	1	0.44 (0.35)
		λ_{34}	1	1		λ_{34}	1	0.79 (0.58)
Method factor 2								
η_5		λ_{45}	1	1	η_5	λ_{45}	1	1
		λ_{55}	1	1		λ_{55}	1	1
		λ_{65}	INV	0.64 (0.08)		λ_{65}	1	1
Latent means		K_1	4.29 (0.07)	3.94 (0.13)		K_1	4.72 (0.09)	4.28 (0.16)

Table 10 Measurement invariance test results for Ukrainian and Russian in Ukraine

Cognitive process		Study 4. Ukraine					
		Ukrainian		Russian		Ukrianian	Russian
Scalar invariance							
ξ_1	Parliament	α_1	INV	0.00	Economy	α_1 0.40 (0.47)	0.00
	Legal	α_2	INV	-0.07 (0.15)	Government	α_2 INV	-0.44 (0.28)
	Politicians	α_3	INV	0.02 (0.13)	Democracy	α_3 INV	-0.15 (0.27)
Metricinvariance							
ξ_1	Parliament	γ_{11}	INV	1.00	Economy	γ_{11} 0.69 (0.12)	1.00
	Legal	γ_{21}	INV	0.85 (0.03)	Government	γ_{21} INV	1.30 (0.10)
	Politicians	γ_{31}	INV	0.78 (0.03	Democracy	γ_{31} INV	1.24 (0.09)
Response process							
Question at time 1							
η_1	Parliament	λ_{11}	1	1	Economy	λ_{11} 1	1
η_2	Legal	λ_{22}	1	1	Government	λ_{22} 1	1
η_3	Politicians	λ_{33}	1	1	Democracy	λ_{33} 1	1
Question at time 2							
η_1	Parliament	λ_{41}	INV	0.95 (0.01)	Economy	λ_{41} 1.19 (0.01)	1.13 (0.02)
η_2	Legal	λ_{52}	INV	0.99 (0.01)	Government	λ_{52} 1.03 (0.01)	1.12 (0.01)
η_3	Politicians	λ_{63}	INV	0.99 (0.01)	Democracy	λ_{63} 1.02 (0.01)	1.10 (0.01)
Method factor 1							
η_4		λ_{14}	INV	0.79 (0.09)	η_4	λ_{14} 1	1
		λ_{24}	1	1		λ_{24} 1	1.65 (0.39)
		λ_{34}	1	1		λ_{34} 1	1
Method factor 2							
η_5		λ_{45}	INV	0.79 (0.09)	η_5	λ_{45} 1	1
		λ_{55}	1	1		λ_{55} 1	1
		λ_{65}	1	1		λ_{65} 1	1
Latentmeans		K_1	5.09 (0.10)	3.73 (0.10)		K_1 4.06 (0.12)	2.79 (0.09)

Taken together, the four studies presented in this research support the idea that survey data from different linguistic groups within a country cannot be aggregated without testing for measurement invariance. In addition, we show that it is important to distinguish the response and cognitive processes of the measurement model. The classical test for measurement invariance is too restrictive, it imposes the unnecessary condition that the reaction to the measurement method should be the same across groups. Using an alternative parameterization that allows the distinction between the cognitive part and the response part, we showed that non-invariance at the response level is more frequent that non-invariance at the cognitive level.

For trust in institutions and at the cognitive level, parameters in Belgium, in Switzerland and in Ukraine were fully scalar invariant, whereas they were partially scalar invariant in Estonia. For satisfaction with politics, full scalar invariance was established in Belgium, in Switzerland and in Estonia. Ukraine was partially scalar invariant.

At the response level, Belgium, Estonia and Ukraine are fully scalar invariant for 'trust in institutions' but in Switzerland misspecifications were present in the method factors. The picture changes in the case of 'satisfaction with politics.' All four countries had misspecifications in the method factors and in Ukraine metric invariance was not achieved. Misspecifications in the method

factors imply that respondents across linguistic groups had different reactions to the way the questions were formulated.

3.5 General discussion

In order to derive substantive comparisons across groups using survey data, it is advisable to test for measurement invariance. But tests using large scale comparative surveys are often limited to testing for invariance across countries. The assumption is that they are homogeneous cultural entities. However, the presence of linguistic groups in multilingual countries set up culturally diverse countries. In the present research, we challenge the assumption that cultural groups within a country are invariant and we extend the test for measurement invariance across linguistic groups.

We suggest that an invariance test in multilingual samples should establish whether linguistic groups interpret concepts in the same way, allowing to take into account how the measurement instrument affects the observed responses, that is, a distinction of the cognitive and the response process in a measurement model should be made.

However, this distinction cannot be done using the well established procedure to test for measurement invariance (Meredith, 1993; Steenkamp & Baumgartner, 1998). It has a known flaw: when the null hypothesis of invariance is rejected, the standard model does not provide information about whether (linguistic) groups are using different interpretations of concepts, there are differences in the

response process, or both. This makes the standard procedure very strict because if differences across groups are only present in the response process (as is the case in the studies presented in this paper), comparisons across groups can be done by correcting for measurement error (Saris & Gallhofer, 2014).

We used an alternative parameterization suggested by Saris and Gallhofer (2014) where the cognitive and the response processes are distinguished to avoid that measurement errors confound differences in the equivalence of the understanding of the concepts across groups. To identify a model where method factors are estimated, we used two observed variables for each indicator, other possibilities can be to incorporate information about the reliability of a measurement instrument in the model, for such procedures the reader is referred to the literature in this topic (Alwin, 2007; Saris & Gallhofer, 2007). Results showed that, in general, the measurement models exhibited invariance in the cognitive level, linguistic groups share a same understanding of the concepts asked, but at the response process there were few but significant differences. Therefore, a distinction seems necessary.

The estimated parameters of the response process showed some significant differences in the reaction to the method of the questionnaire across linguistic groups. If they are not taken into account, these variations disturb the standard test for measurement invariance making it more likely to be rejected. In the four studies included in this paper, response differences across groups were

more common than differences in the interpretations of the concepts.

The results have two implications for survey research. The first is that in multilingual samples with questionnaires in more than one language, invariance should be tested across linguistic groups before aggregating data at the country level.

Asecond implication is that when there are differences in the response process, comparison of relationships and means across countries are affected by measurement error. Estimating means and relationships using latent models solve this problem, but composite scores calculated on the basis of observed variables should not be used. As the response process presents significant differences across groups, if composite scores are estimated directly from observed variables, measurement error will affect substantive conclusions.

Appendix 3.1 Global fit indices of the models

Belgium, Trust in institutions: DF = 19; χ^2= 52.18 (p = 0.00); RMSEA = 0.056, 90 % CI for RMSEA = (0.038 ; 0.074); CFI = 1.00.

Belgium, Political satisfaction: DF = 19; χ^2= 25.25 (p = 0.15); RMSEA = 0.031, 90 % CI for RMSEA = (0.0 ; 0.064); CFI = 1.00.

Estonia, Trust in institutions: DF = 18; χ^2= 70.87 (p = 0.00); RMSEA = 0.065, 90 % CI for RMSEA = (0.049 ; 0.082); CFI = 1.00.

Estonia, Political satisfaction: DF = 18; χ^2= 29.86 (p = 0.039); RMSEA = 0.046, 90 % CI for RMSEA = (0.0 ; 0.076); CFI = 1.00.

Switzerland, Trust in institutions: DF = 16; χ^2= 59.49 (p = 0.00); RMSEA = 0.044, 90 % CI for RMSEA = (0.032 ; 0.056); CFI = 1.00.

Switzerland, Political satisfaction: DF = 18; χ^2= 36.32 (p = 0.0064); RMSEA = 0.056, 90 % CI for RMSEA = (0.029 ; 0.082); CFI = 1.00.

Ukraine, Trust in institutions: DF = 19; χ^2= 36.81 (p = 0.0084); RMSEA = 0.038, 90 % CI for RMSEA = (0.019 ; 0.056); CFI = 1.00.

Ukraine, Political satisfaction: DF = 17 ; χ^2= 56.73 (p = 0.00); RMSEA = 0.100, 90 % CI for RMSEA = (0.072 ; 0.13); CFI = 0.99.

Chapter 4

Exploring language effects in cross-cultural survey research: Does the language of administration affect answers about politics?

4. Exploring language effects in cross-cultural survey research: Does the language of administration affect answers about politics?[10]

Abstract

We study if the language of administration of a survey has an effect in the answers of bilingual respondents to questions measuring political dimensions. This is done in two steps. In the first we test whether the measurement instruments are equivalent for a same individual in two languages. After measurement invariance is established, we tested if latent mean differences are significant across the two languages. We also test if the correlation of a same concept in two languages is equal to one or not. Results show evidence for language effects, the latent correlation was below one, although mean differences were not significant. We use data of the LISS migration panel in a within subject design: respondents answered a questionnaire in Dutch first and then in their (second) native language amongst Arabic, English, German, Papiamento and Turkish.

4.1 Introduction

Populations of interest in large scale cross-national survey projects are linguistically diverse. For instance, in the European Social Survey (ESS) and in the European Values Survey (EVS) questionnaires are translated when at least 5% of the population is

[10] Submitted for publication and under review at the time of the dissertation's defense.

native speaker of a language (Dorer, 2012; European Values Survey, 2010), but little is known about the consequences of this decision. In the present research, we study if the language of administration of a survey predicts *bilingual* respondents' answers to questions measuring *political dimensions*. We define bilingual individuals in terms of language use, that is, individuals who have the ability to write, speak, read, and listen in two languages. Furthermore, they use both languages in their daily life: in their main activities such as work or school and with their friends and relatives (Grosjean, 2014).

Language effects can emerge in comparative survey research because problematic translations fail to reproduce the same stimuli across languages (Pennell, Harkness, Levenstein, & Quaglia, 2010). Secondly, when the language of an interview activates cultural orientations driving individuals' responses (Luna et al., 2008). Language is a strong cultural carrier (Cohen, 2009) and bilingual individuals tend to live in mixed cultural environments. Cultural orientations may influence thoughts, cognitions and behaviour (Oyserman & Lee, 2008) and this in turn may affect the way respondents interpret and answer survey questions. Although translation issues have gained importance in comparative survey methodology (Janet A. Harkness, Villar, et al., 2010), the effects of the language of administration in the responses to a questionnaire have received little attention in survey research.

Research about language effects in the answers to measurement instruments has been conducted mainly in the fields of sociocultural psychology and psycholinguistics. Although diverse in methods and approaches, it has consistently been found that the language of administration of a questionnaire has an effect on the answers that bilingual individuals give to cultural and self-identity items (cf. S. X. Chen & Bond, 2010). If the language of the questionnaire influences bilinguals' responses, cross-cultural differences could be confounded to some degree regardless of respondents' true opinions. Moreover, as the proportion of bilingual individuals is different across countries, the potential impact of this bias in a cross-national survey is unknown.

In the present research, we tested for language effects in political dimensions ruling out translation issues. We conducted a within-subject study of bilinguals representing five minority groups in the Netherlands, a country with high linguistic diversity. Participants answered a questionnaire in Dutch and in their (other) native tongue: Arabic, English, German, Papiamento or Turkish. The first step was to test for measurement equivalence. Once equivalence was established, we tested whether the correlation of a concept in two languages is equal to one. Thirdly, we tested if differences in latent means across languages were significant. The article proceeds as follows: In the first part, we make a more in-depth introduction of the mechanisms behind the effects of the language of administration in the answers to measurement instruments. In the next section, we introduce the operationalization of the concepts

and the models used to test for language effects: 'Trust and need for change in institutions' and 'Satisfaction and need for change in politics and the economy'. In the following section, we explain our methodology: the procedures regarding the estimation and testing of the models. Next, we present the survey data in which the study is embedded. Finally, we summarize the results and discuss the general findings.

4.2 Language effects in responses to measurement instruments

The mechanism behind the adaptation of responses as a function of the language in an interview can be explained by the theoretical frameworks of *acculturation* (Schwartz et al., 2014) and *cultural frame switching* (*CFS,* Hong et al. 2000). A bilingual personmay gradually develop into a bicultural person (Grosjean, 2014). As language is a strong cultural carrier (Cohen, 2009), individuals who master two languages may start an acculturation process internalizing to some extent the cultural attitudes and values attributable to the second language (Bond & Yang, 1982). Acculturation operates in three dimensions. The first is at the level of social behaviours or *practices*, such as cuisine preferences, language use and the choice of friends. The second is the acquisition of cultural *values*, for instance the importance of individualism versus collectivism. The third dimension is regarding identification: the attachment to a cultural, ethnic or national group (Schwartz et al., 2010).

Bicultural individuals have internalized two sets of culturalorientations i.e. cultural systems, even with conflicting premises.They can be activated independently by relevant cultural stimuli. CFS takes place when a person uses one cultural system instead of the other to react to social cognitions. For this to happen, cultural orientations are activated and they become highly accessible in the mind of the person. In an experimental design, Hong et al. (2000) presented bicultural subjects with iconic images of Chinese or American cultures. Participants were asked to interpret an ambiguous social scene of a fish bank. It was not clear if the stimulus represented a social event or not. It was not evident whether it was an individual or a group action. Participants used the cultural frame that was recently activated by the cultural symbols to interpret the ambiguous stimulus. They provided more group-oriented answers when they were primed with Chinese culture icons. On the other hand, they gave more individual-centred answers when the American icons were shown first. Research has shown that the language of the interview can be a powerful activator of cultural-specific mind-sets in bilinguals and in consequence, individuals' answers to a questionnaire are adjusted (Bond & Yang, 1982; S. X. Chen, Benet-Martínez, & Ng, 2014; S. X. Chen & Bond, 2010; Luna et al., 2008; Schwartz et al., 2014, 2010; Yang & Bond, 1980).

Previous research about language effects has several limitations. Primarily, it has been done with Asian subjects from Hong Kong (HK) comparing their responses in Chinese and English languages,

followedby research on the differences between Spanish-English languages in Hispanic communities in the United States. However, the dichotomies Chinese-Westerner or Hispanic-Westerner (where Western means English language or American culture) may be very specific cases. Both Chinese and Hispanic cultures have emphasis oncollectivism as an archetypal trait, whereas preference for individualism is regarded as a Western archetype (Yoon, 2010). Respondents from highly communitarian cultures are more sensitized to contextual clues. They may assume that a certain type of culturally-oriented response is expected (Lechuga, 2008). Moreover,evidence shows that the distance between Asian cultures and Western culture is perceived as very large (Minkov, 2007). Other languages have been explored in few cases: Arabic (Botha, 1968), Afrikaans (Botha, 1970), Cebuano (Watkins & Gerong, 1999), French (Botha, 1968; Candell & Hulin, 1986), Greek (Richard & Toffoli, 2009; Triandis, Davis, Vassiliou, & Nassiakou, 1965), Korean (Perunovic, Wei, Heller, & Rafaeli, 2007) and Russian (Marian & Neisser, 2000) and only one large scale research design was conducted in more than twenty languages (Harzing, 2006). When language effects have been tested in other cultural contexts, findings have not been replicated completely. To what extentcanlanguage effects be generalized to individuals of cultural backgrounds that are not Chinese or Hispanic?

Secondly, the state of the art suggests that only answers to concepts about cultural and self-relevant identity domains are affected by the language of the measurement instrument. Luna et al. (2008) states

that CFS only happens in bicultural bilinguals. The information that monocultural bilinguals have associated to their second language is not related to self-relevant identity constructs; it does not affect how they see themselves. Several studies have found that language effects are mediated by individual characteristics related to *biculturalism* e.g. exposure to both cultures and the extent to which they are perceived as compatible or oppositional to *language acquisition* e.g. learning languages in different settings and time of first exposure to each language (Benet-Martínez & Haritatos, 2005; Benet-Martinez et al., 2002; Dixon, 2007; Ji et al., 2004; Ross, Xun, & Wilson, 2002; Tyson, Doctor, & Mentis, 1988).

Benet-Martínez, Lee, and Leu (2006) proposed that bicultural thinking about culture is more sophisticated than that of monocultural individuals. They are more experienced in dealing with cultural informationbecause of their frequent CFS experiences. As a consequence, biculturals would have more complex cultural representations than monoculturals, but this trend is not expected in culturally neutral domains. However, with the exception of physical and mental health for which language effects did not emerge (Elliott, Edwards, Klein, & Heller, 2012; Peytcheva, 2008), *culturally neutral topics* have been tested in few cases. Language effects have been studied in laboratory-settings where culturally neutral topics were far too neutral and of no relevance to social or political dimensions e.g. geometric figures or landscapes.

Language effects have been found consistently in responses to questionnaires about cultural dimensions (Benet-Martínez et al., 2006; Bond & Yang, 1982; Harzing, 2005; Lechuga, 2008; Schwartz et al., 2014; Toffoli & Laroche, 2002; Triandis et al., 1965; Yang & Bond, 1980), personality perceptions (S. X. Chen et al., 2014; S. X. Chen & Bond, 2010; Ramírez-Esparza et al., 2006), feelings (Marian & Kaushanskaya, 2004; Perunovic et al., 2007), autobiographical memory (Marian & Neisser, 2000; Schrauf & Rubin, 2000), subjective evaluative ratings (Bond, 1985; Elliott et al., 2012; Pierson & Bond, 1982; Toffoli & Laroche, 2002) and self-relevant identity constructs (Dixon, 2007; Kemmelmeier & Cheng, 2004; Pierson & Bond, 1982; Ross et al., 2002; Trafimow, Silverman, Fan, & Fun Law, 1997). Topics in which language effects emerged have, in many cases, not only a cultural component but also an emotional or highly personal one.

A third limitation in the study of language effects in the answers to questionnaires is of a methodological nature. Most published work has tested for language effects by mean differences in composite scores. There are several problems with this approach. The first is the implicit assumption that the measures are statistically equivalent across linguistic groups. Measurement equivalence is a prerequisite for cross-cultural comparison of models, relationships and means (cf. Davidov et al., 2014; Meredith, 1993; Vandenberg & Lance, 2000; Vandenberg, 2002). In other words, before interpreting differences in responses, it is essential to test if the same

measurement model on therelationship between indicators and latent variables holds in both languages.

Figure 11 illustrates a case where the measurement instruments are equivalent across linguistic groups. The linear function that links the observed responses to the latent opinion is the same in language A and language B. In contrast, Figure 12 shows a case when the response functions are different for each linguistic group.They have the same score on the latent variable (opinion) on a certain topic but they differ in the intercept, τ_i, and/or in the slope, λ_i, therefore, they differ in their score given as a response, y. In an invariant situation, both groups expressed their opinion in the same way. In a non-invariant situation, Group A expressed its opinion in a very extreme way, while B did the opposite. This latter result would not allow for the comparison of relationships and means across languages because each of them had given different answers for the same opinion.

Figure 11. Measurement equivalence

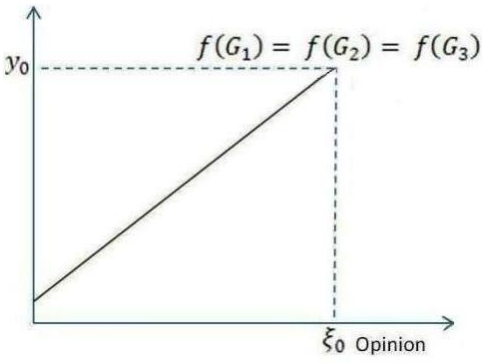
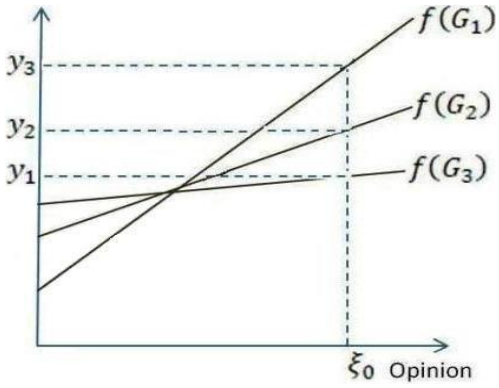


Figure 12 Measurement equivalence is not established



With some exceptions, measurement invariance has not been established prior to test for language effects in bilingual individuals (Candell & Hulin, 1986; Richard & Toffoli, 2009; Schwartz et al., 2014 test for measurement invariance and language effects).

A second methodological problem in the analysis of language effectsis that manifest variables are not measurement-error free. When differences in observed means have not been found significant, the conclusion has been that language effects are negligible. Only when full invariance is found can composite scores be used directly, but they should be corrected for measurement error to derive unbiased results (Saris & Gallhofer, 2014). When partial invariance is found (Byrne, Shavelson, & Muthén, 1989), that is, when at least two of the measures of a concept are invariant, latent means should be used because composite scores are not adequate.

A third problem is that by comparing means based on composite scores or latent means, it is not tested whether the conceptual

associations that individuals retrieved when they use one language or the other are the same. Equal latent means do not imply that the correlation between a latent concept in one and the other language is one. Richard and Toffoli (2009) found that although the factorial structure of a construct (configural invariance) and the way respondents answered (factor loadings invariance) were the same in two languages, the covariances between the latent variables were significantly different across Greek and English. They argued that respondents had different conceptual associations in each language.

Evidence suggests that bilinguals may use different conceptual associations in each language, except in the cases where an exact translation exists (Ji et al., 2004; Luna et al., 2008). For instance, the language of an interview has been found to be a powerful activator of memories. Marian and Neisser (2000) and Schrauf and Rubin (2000) found that participants retrieved auto-biographical experiences associated to the use of one language consistent with the language of the interview. When respondents were interviewed in Russian (English), they remembered more experiences of their Russian-speaking (English-speaking) period of their lives (Marian & Neisser, 2000). In Hispanic bilinguals, autobiographical memories were encoded and retrieved in Spanish for events associated to the use of Spanish language, and in English for events in which English language was used (Schrauf & Rubin, 2000).

In summary, in the study of language effects, statistical equivalence across languages remains empirically unexplored. Furthermore,

when statistical equivalence is established, a test where latent (or observed) mean differences are not significant does not rule out the possibility of language effects. It indicates that the distribution of the variable in the two languages is the same (equality in the location parameter) but that respondents can still have different conceptual associations in each language.

We propose a different approach to test for language effects. We usea specific application of a LISREL model (Jöreskog & Van Thillo, 1973), which we call in the following sections the *baseline model*. This model assumes linear relationships between indicators (observed variables) and unmeasured constructs (latent variables). In this model, the language specific measurement error component is taken into account when the relationship between the indicators and the latent variables is specified. In the first step of this model, we test whether the relationship across indicators and latent variables is the same in both languages. This is the test for measurement invariance.

Once it is established that the measurement model is equivalent, we are able to test structural relationships of latent variables in two languages. We test whether two latent variables represent the same variable of interest by testing if its correlation is equal to one (Jöreskog, 1971; Saris, 1982a, 1982b). In other words, if two latent variables are the same. Having corrected for measurement error, a very high correlation, say 0.90 would imply that the variables are

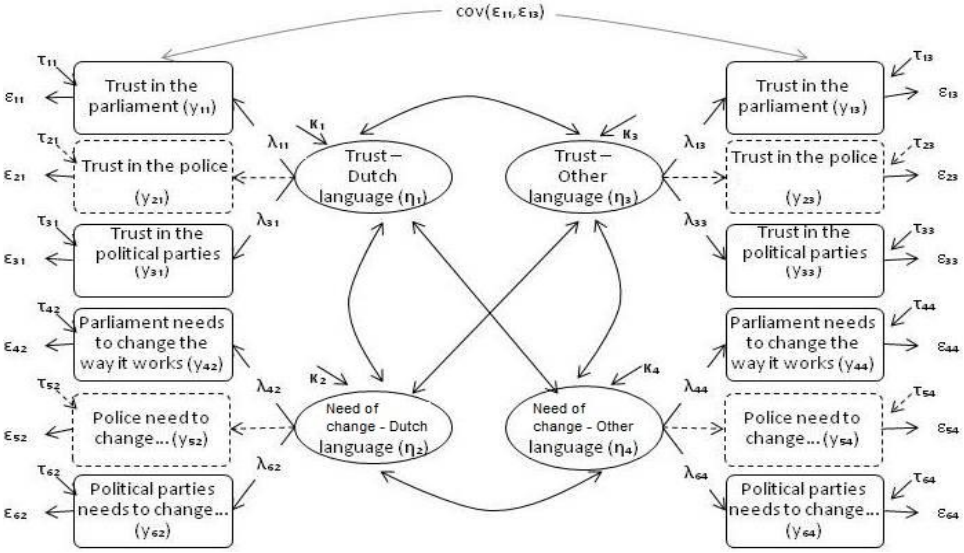
very similar across languages, but that they nevertheless have a unique component that causes them to be not exactly the same.

4.3 The concepts of interest: Political satisfaction and trust in institutions

The study of trust in institutions and political satisfaction has a long tradition in Political Science (Easton, 1975; Inglehart, 1977; Kaase, Newton, & Scarbrough, 1997; Levi & Stoker, 2000). In advanced democracies, it is unlikely that people reject the idea of democracy when satisfaction with politics and trust in the institutions is poor. However, persistent dissatisfaction and distrust increase citizens' concerns about the functioning of democracy and their perception that a systemic change is needed (Brons, 2014; Easton, 1975; Hendriks, 2009; Inglehart, 1977). We test whether the answers that respondents give to the key dimensions of 'trust in institutions' and 'political satisfaction' vary as a function of the survey language. For these concepts we use a similar operationalization previously asked in cross-national surveys (European Social Survey, 2015a). In addition, we develop a measure of respondents' perception of political change. We operationalize the concept 'political change' in a survey questionnaire following the three step procedure for formulating survey questions suggested by Saris and Gallhofer (2014). Appendix 4.1 shows the development of the measures used in Model 1 and Model 2.

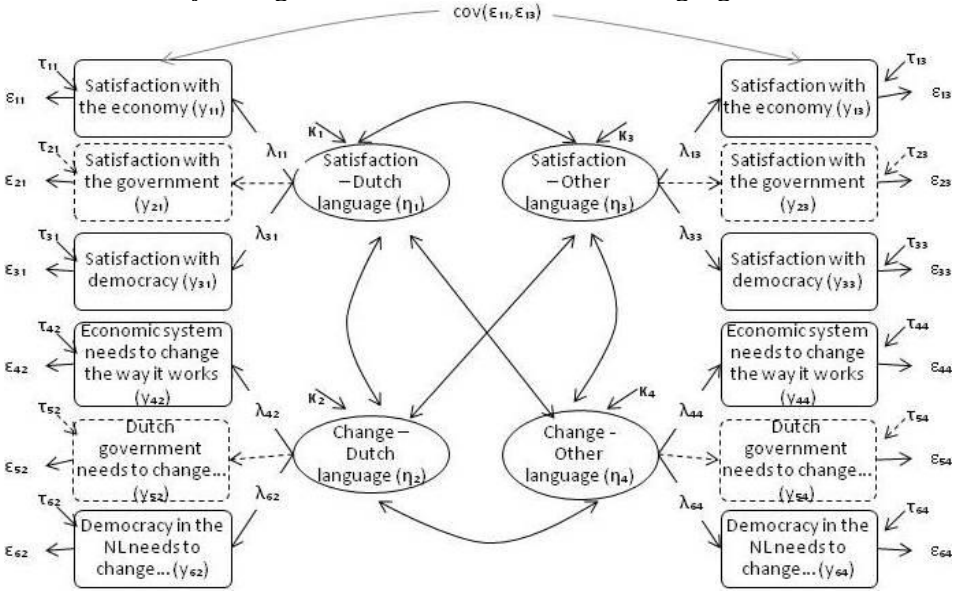
The first model we test,'Trust and need of change in institutions' (Model 1), consists of two concepts-by-postulation (CP) -or latent concepts. The first labelled 'Trust'in institutions can be decomposed into simpler concepts -or concepts-by-intuition (CI) directly measured by survey questions: 1) trust in the parliament, 2) trust in the political parties and 3) trust in the police. The second CP, 'Need for change' consists of three CI representing evaluative beliefs about the need for change in the way the aforementioned institutions work. In the same way, 'Satisfaction and need of change in politics and the economy' (Model 2) includes two CP: 'Satisfaction' decomposed into three CI: a feeling of satisfaction with the economy; satisfaction with the government, and satisfaction with democracy in the Netherlands. 'Need for change'is also made up of three evaluative beliefs. The need for change in the economy; the need for change in the way democracy works and the need for change in the government.

Figure 13 Basic factor structure of Model 1. Baseline model for 'Trust and change in Institutions' Bilinguals in Dutch and their second language



Note: We introduce $cov(\varepsilon_{11}, \varepsilon_{13}), cov(\varepsilon_{21}, \varepsilon_{23}), \ldots, cov(\varepsilon_{52}, \varepsilon_{54}), cov(\varepsilon_{62}, \varepsilon_{64})$. They are shown in the figure only once. (Error) variances are not shown for simplicity. The dotted line is for a loading parameter constrained to 1 and an intercept to zero to fix the scale of the latent construct (identification of the model).

Figure 14 Basic factor structure of Model 2. Baseline model for'Politics and the Economy' Bilinguals in Dutch and their second language



Note: We introduce $cov(\varepsilon_{11}, \varepsilon_{13}), cov(\varepsilon_{21}, \varepsilon_{23}), \ldots, cov(\varepsilon_{52}, \varepsilon_{54}), cov(\varepsilon_{62}, \varepsilon_{64})$. They are shown in the figure only once. (Error) variances are not shown for simplicity. The dotted line is for a loading parameter constrained to 1 and an intercept to zero to fix the scale of the latent construct (identification of the model).

The models to be tested are presented in Figure 13 and Figure 14. In the figures, the left hand side represents the model using the answers from the Dutch questionnaire. On the right hand side, the model corresponds to the same individuals answering in a second language (Arabic, English, German, Papiamento or Turkish).

The η_j represents the jth latent CP; the y_{ij} is the ith observed variable (CI) for the jth latent trait and ε_{ij} are the disturbance terms; the λ_{ij} are the loadings; τ_{ij} are the intercepts and κ_j the latent means. It is assumed that the disturbance terms have a mean of zero and that they are uncorrelated with the latent variables. The disturbance terms are a combination of random errors and unique components. Thus, the unique components are correlated for the same observed variable in different languages denoted by $cov(\varepsilon_{11}, \varepsilon_{13}), cov(\varepsilon_{21}, \varepsilon_{23}), \ldots, cov(\varepsilon_{52}, \varepsilon_{54}), cov(\varepsilon_{62}, \varepsilon_{64})$. The other disturbance terms are assumed to be uncorrelated.

The latent variables (η_j) are correlated with each other. In order to assign a scale to the latent CP, for each one, the loading of one observed variable is fixed to one, and the respective intercepts to zero (depicted with a dotted line in the picture).

4.4 Method

We tested for measurement equivalence within subjects in two languages through confirmatory factor analysis (CFA) (cf. Davidov et al., 2014; Horn & McArdle, 1992; Meredith, 1993; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000)[11]. We tested a series of models starting with the baseline models shown in Figure 13 and Figure 14 and introducing consecutive equality constraints in the parameters. Firstly, we tested that the same configuration of the factorial structure held in both languages. Secondly, the configural model was restricted to one where the factor loadings were constrained to be equal for the same manifest variable in a different language ($\lambda_{11} = \lambda_{13}; \lambda_{31} = \lambda_{33};\ \lambda_{42} = \lambda_{44};\ \lambda_{62} = \lambda_{64}$). When this restriction is not rejected, it is implied that comparisons of relationships across groups can be made. Thirdly, in addition to equivalence in the factor loadings, the intercepts were constrained to be equal ($\tau_{11} = \tau_{13};\ \tau_{31} = \tau_{33};\ \tau_{42} = \tau_{44};\ \tau_{62} = \tau_{64}$). When the restriction in the intercepts is not rejected, it is implied that comparisons of means can also be made across languages.

Once equivalence in the measurement parameters was established we further constrained the models to test 1) whether the correlation between a construct in Dutch and in another language is equal to one ($\rho(\eta_1, \eta_3) = 1;\ \rho(\eta_2, \eta_4) = 1$). Failing this test is interpreted

[11] We estimated the models using Maximum likelihood estimation with 'lavaan' package for structural equation modeling (SEM) (Rosseel, 2012) in R 3.1.2 statistical environment (R Core Team, 2015). A repository with the scripts and the data used in the analyses in this paper is found at this link: https://github.com/dianazr/LanguageEffectsCrossCulturalSurvey

as the variables "reflect[ing] differences in conceptual associations among the true scores" (Vandenberg, 2002, p. 142) and that they are not exactly the same because they have a unique component in each language (Saris, 1982a). We also test 2) for invariance in the factor means ($\kappa_1 = \kappa_3$; $\kappa_2 = \kappa_4$). This restriction tests for differences between the two languages in the mean (latent) scores of the constructs of interest.

a) Estimation and testing of the models

We used the likelihood ratio test (LRT) in combination with the Judgement Rule (JRule) approach to test our models (Saris et al., 2009)[12]. Goodness of fit (GoF) indices of structural equation modelling (SEM) are quite controversial (Cheung & Rensvold, 2002). Commonly used fit criteria such as the Chi-square and the Root Mean Squared Error of Approximation (RMSEA) do not control for Type II error. Saris et al. (2009) suggest that when GoF measures are used, a misspecified model can be accepted whereas a model with irrelevant misspecifications can be rejected. Moreover, GoF measures do not give an insight into which elements in a model are misspecified. A misspecification occurs if a parameter has been given a fixed or constrained value, which is incorrect in the population of study (Hu & Bentler, 1998). The difference in the LRT indicated whether the GoF is significantly worse for progressively more restrictive models. The JRule approach (Saris et al., 2009) identifies whether fixed or constrained parameters are

[12] Global fit indexes are reported in Appendix 4.3

misspecified at each level of the equivalence tests specified in the previous section. In this way, we tested directly for misspecifications in the model while taking into account the power of the test for each misspecification. JRule works by combining knowledge of: (a) the size of the misspecification (Expected Parameter Change, EPC); (b) the modification index (MI), its impact on the fit if the parameter was freed in the model; and (c) the power of the test in detecting the misspecification[13]. Table 11 shows the decision rules to different situations based on the MI and the power of the test.

Table 11. The decisions to be made in the different situations defined on the basis of the size of the modification index (MI) and the power of the test.

	High power	Low power
Significant MI	See whether or not the size of the EPC is larger than the threshold. Parameter set free if it is	Misspecification present (Parameter is freed)
Non-significant MI	No misspecification (Parameter is not freed)	Inconclusive (Parameter is not freed)

Saris et al. (2009) proposed a heuristic to choose the threshold for relevant differences. Following this recommendation, we chose a power of 0.8 to detect standardized loading differences larger than 0.1 and intercept differences larger than 5% the length of the scales (all measures had 11-point scales). If a constrained parameter was misspecified according to JRule, it was freed and the null

[13] The JRule approach for R was programmed as the 'miPowerFit' function in the semTools package (Pornprasertmanit, Miller, Schoemann, & Rosseel, 2014). The 'miPowerFit' function takes (a) and (b) from the SEM output of 'lavaan' and estimates (c).

hypothesis of invariance in that restriction rejected. Once measurement equivalence was established, we set a threshold of 0.55 for differences in standardized latent means, which equals 5% of the items' scale, and a power of the test of 0.80. The same procedure was used to test for equality of latent covariances/correlations. We restricted them to be equal between groups and tested whether this restriction was misspecified or not using a power of 0.80 and a threshold of 0.10 for differences (see Cieciuch et al., 2015; Van der Veld & Saris, 2011 for applications of Jrule as a model testing method to test for mesurement invariance).

4.5 Data

a) Participants

The empirical findings in this article are based on a two wave study conducted between April and June, 2013 at the Measurement and Experimentation in the Social Sciences (MESS) Immigrant Panel administered by CentERdata at Tilburg University, The Netherlands. The Immigrant panel was a probability-based online project in which researchers could submit proposals for fieldwork at no cost. Respondents were recruited based on stratified sampling using the population registry as sampling frame. Participants have foreign backgrounds of four major migration groups in the Netherlands (first and second generations of western and non-western origin). They were provided with internet and a laptop to

answer monthly surveys and received an economic incentive for each completed questionnaire.

b) Data collection

The objective of Wave 1 was to select the languages in which translations would be obtained to test for language effects in a within-subject design in Wave 2. Wave 1 included 989 bilingual participants. They mentioned 74 languages as their native tongues. We selected the five languages in which respondents had the highest self-reported proficiency and the group was of at least 30 individuals: Arabic, English, German, Papiamento and Turkish. The source questionnaire was developed simultaneously in Dutch and English, translations into the other four target languages were done according to a committee approach with two independent translators and an adjudicator that harmonized differences, questions were pretested with at least one person in each language. This approach was based on the TRAPD (Translation, Review, Adjudication, Pretesting and Documentation) procedure which is the gold-standard approach in survey translation (Janet A. Harkness et al., 2004).

In the second wave, the questionnaire was presented to 308 bilingual panel members, and it was fully completed by 255 respondents (83%). Due to the small number of individuals per language, the analysis was done by linguistic group. The results presented in the next section are derived from this final sample size. Table 12 shows the composition of the sample in Wave 2 according to language and completion rates.

Table 12. Completion rates by linguistic group in Wave 2

Language	Selection	Complete	Completion rate
English	126	104	82.5%
Papiamento	36	31	86.1%
Arabic	36	30	83.3%
German	38	35	92.1%
Turkish	72	55	76.4%
Total	308	225	85.5%

The questionnaire in Wave 1 included questions in Dutch about language use and knowledge and core questions (described in Section 4.4). In the first set, all participants self-rated their ability in writing, listening, speaking and reading Dutch and their (second) native language in an 11-point scale (from 0 to 10). Table 13 shows the mean and standard deviation of self-reported proficiency in both languages. Results are shown only for participants who later on participated in Wave 2.

Table 13.Self-reported proficiency in Dutch and target languages

Language group	Mean self-reported proficiency in Dutch (Standard deviation)				Mean self-reported proficiency in target language (Standard deviation)			
	Write	Read	Speak	Listen	Write	Read	Speak	Listen
English (n=104)	7.6 (2.4)	9.0 (1.4)	8.8 (1.5)	9.0 (1.5)	8.7 (1.7)	9.1 (1.4)	9.1 (1.2)	9.3 (1.3)
Papiamento (n=31)	7.1 (2.7)	8.5 (2.3)	8.6 (1.3)	8.9 (1.3)	6.3 (3.1)	7.4 (2.7)	8.5 (2.2)	8.8 (2.1)
Arabic (n=30)	5.9 (2.4)	7.0 (2.5)	7.0 (2.5)	7.4 (2.4)	7.8 (2.6)	8.2 (2.5)	8.8 (2.1)	9.0 (1.9)
German (n=35)	8.0 (1.8)	9.6 (0.8)	9.2 (1.3)	9.7 (0.7)	7.4 (2.4)	9.1 (1.3)	8.3 (2.1)	9.3 (1.1)
Turkish (n=55)	7.1 (2.5)	8.0 (2.2)	7.8 (2.1)	8.1 (2.0)	7.4 (2.5)	7.3 (2.6)	7.8 (2.2)	8.0 (2.0)

Moreover, as shown in Table 14, bilingual participants were asked which language they use (d) more frequently at home, at school or work, with friends and with their parents. Proportions indicate that bilingual participants live in a highly mixed cultural environment, combining their languages in different aspects of life. Linguistic minorities represented in this study have a large usage of the Dutch language in several contexts. However, they use their (other) native tongue in personal contexts such as at home and with their parents, at school and work, their predominant daily language is Dutch. This is also the case for language usage with friends except for Turkish participants, whose usage is more balanced in both languages. Among German speakers, German language is less frequent in all aspects of life except with their parents.

Table 14.Self-reported language useinDutch and target languages

Language group	Dutch language most frequently used... (%)				Target language most frequently used... (%)			
	At work/school	With friends	At home	With parents	At work/school	With friends	At home	With parents
Arabic	92.6	56.7	40	0	3.7	33.3	53.3	88.2
English	70.2	81.7	51.9	43	29.8	16.3	47.1	70.7
German	85.7	97.1	85.7	26	8.6	2.9	11.4	47.3
Papiamento	100	70.9	54.7	14.2	--	25.5	45.4	71.2
Turkish	90.2	45.5	21.8	6	7.8	49.1	69.1	88

Note: When percentages adding Dutch and target language do not add 100, 'other' language was reported as most used.

Wave 2 consisted of three parts. In the first, individuals answered the core questions in Dutch. Then, in the second part, they answered an unrelated questionnaire about different topics such as ideal body types, nature preservation, and King's Willem-Alexander succession. In the third part, they answered the core questions in Arabic, English, German, Papiamento or Turkish depending of the information they provided in the first wave. Although memory effects are a possibility, they can be controlled for repetitions in survey interviews when other questions are asked in between (Van Meurs & Saris, 1990).

4.6 Results

In this section we describe the results at each level of invariance testing. We combine the LRT and the JRule approach to evaluate the baseline models versus progressively more restricted models to test for measurement equivalence.

a) Equivalence in the factorial structure

Following the JRule test of local misspecifications, the baseline Model 1 and Model 2 were slightly modified. The p-value of the LRT is significant for the fit of the baseline model versus a model with some correlated errors as shown by Table 15. In Model 1 (Figure 13), we introduced two error covariances. The first was across the disturbance terms of the observed variable ‘trust in the police’ and ‘need for change in the way the police works’

($cov(\varepsilon_{21}, \varepsilon_{52}) = cov(\varepsilon_{23}, \varepsilon_{54})$) and the second between "trust in political parties' and 'need for change in the political parties' ($cov(\varepsilon_{31}, \varepsilon_{62}) = cov(\varepsilon_{33}, \varepsilon_{64})$). Both correlations are constrained to be equal across languages. In Model 2 (Figure 14), we introduced three error covariances restricted to be equal between languages: 1)'satisfaction with the economy' and 'need for change in the economy' $cov(\varepsilon_{11}, \varepsilon_{42}) = cov(\varepsilon_{13}, \varepsilon_{44})$, 'satisfaction with the government' and 'need forchange in the government' $cov(\varepsilon_{21}, \varepsilon_{52}) = cov(\varepsilon_{23}, \varepsilon_{54})$ and 'satisfaction with the way democracy works in the NL' and 'change in the way democracy works in the NL' $cov(\varepsilon_{31}, \varepsilon_{33}) = cov(\varepsilon_{62}, \varepsilon_{64})$. Correlated errors improved the fit of the model and are constrained to be equal across languages. Configural invariance is established because the same linear relationships exist between the indicators and the latent variables in both languages.

b) Equivalence in the factor loadings

Once we established configural equivalence, we constrained the corresponding factor loadings to be equal across languages. As shown in Table 5, the LRT of the configural Model 1 and Model 2 were not significantly different than the restricted models. According to JRule, this restriction was not misspecified. Therefore, equivalence in the factor loadings was established in both models. This test indicated that the strength of the relationship between the observed variables y_{ij} and the underlying jth latent variable is the same in both languages.

c) Equivalence in the intercepts associating manifest and latent variables across languages.

There were no significant misspecifications in the intercepts thus the models are scalar invariant. Furthermore, the LRT did not show that the fit was different between a model constraining loadings and a more restricting one which constrains intercepts. Full measurement equivalence was established in Model 1 and Model 2.

**Table 15 Likelihood ratio test - Baseline versus restricted models –
Within subject measurement equivalence in Dutch and a second language**

	Model 1. Trust and change in institutions					Model 2. Fairness, satisfaction and change in politics and the economy				
	DF	χ^2	$\Delta\chi^2$	Δ DF	$P(>)\chi^2$	DF	χ^2	$\Delta\chi^2$	Δ DF	$P(>)\chi^2$
Baseline model	42	209.94				42	232.8			
Baseline model + correlated errors	40	158.90	51.04	2	<0.001***	39	172.4	60.42	3	<0.001***
Invariance of loadings	44	164.96	6.06	4	0.19	43	175.4	2.996	4	0.558
Invariance of intercepts	48	170.78	5.82	4	0.21	47	179.8	4.415	4	0.353
Significance codes: 0.001> '***'; 0.01 > '**'; 0.05 > '*'; 0.1 > '·'										

d) Within-subject structural equivalence in two languages

- Test for cross-correlations equal to one

We tested whether the correlations between a latent variable in Dutch and the same latent variable in another language was equal to one, $\rho(\eta_1, \eta_3) = 1;\ \rho(\eta_2, \eta_4) = 1$)[14]. This was not the case in either Model 1 or in Model 2. Both the LRT and JRule indicated that this restriction should be rejected (Table 16). In Model 1, the correlation between 'trust' in Dutch and 'trust' in a second language was 0.78 ($\rho(\eta_1, \eta_3)$); and 0.64 between 'change' in Dutch and 'change' in a second language ($\rho(\eta_2, \eta_4)$). In Model 2, The correlation between the construct for 'satisfaction' in Dutch and 'satisfaction' in another language is not equal to one, but significantly lower (0.79) ($\rho(\eta_1, \eta_3)$). In the case of the CP 'change', the correlation between Dutch and a second language was of 0.71 ($\rho(\eta_2, \eta_4)$).

- Test for equal factor means

The LRT of the Model 1 and Model 2 restricting latent means are indicated in Table 16. In Model 1, the LRT indicated that the fit of the model was not significantly different from the one restricting intercepts. In addition, according to JRulewe did not find

[14] To estimate latent correlations and test whether or not they were one, two additional restrictions were imposed to the scalar models: the first was to fix the variances of the latent variables to one. The second, fixing the latent covariances of the same concepts in different languages to one. Using these constraints, the model estimates the matrix of standardized latent covariances, which are the latent correlations.

misspecifications in the equality constraints of the latent means. In Model 2, the LRT indicated that the fit of the model restricting latent means was significantly worse than the one which estimated the means without constraints. However, at the threshold level of 0.55 (5% of an 11-point scale), JRule did not show any significant differences inlatent mean differences. By decreasing the threshold to detect deviation to 0.15 with a power of 0.80, JRule indicated that the equality constraints $\kappa_1 = \kappa_3$ and $\kappa_2 = \kappa_4$ were misspecified. The unstandardized estimate for the factor mean of 'satisfaction'was of 3.61 (se = 0.13) in Dutch language (κ_1) and 3.87 (se = 0.12) in the second language (κ_3). The unstandardized latent mean of 'change' was 6.98 (se = 0.12) in Dutch (κ_2) and 6.81 (se = 0.12) in the respondents' second language (κ_4). This result indicates that the mean scores of the underlying constructs that build Model 1 are significantly different in Dutch and in a second language for the same individual, however the difference is estimated around 1.5%. It is quite smaller than the threshold for mean differences established in Section 4.4.

Table 16. Likelihood ratio test - Within subject differences in latent means and covariances

	Model 1. Trust and change in institutions					Model 2. Fairness, satisfaction and change in politics and the economy				
	DF	χ^2	$\Delta\chi^2$	Δ DF	$P(>)\chi^2$	DF	χ^2	$\Delta\chi^2$	Δ DF	$P(>)\chi^2$
Invariance of intercepts	48	170.78				47	179.83			
Correlations test[a]	54	417.32	246.54	6	<0.001***	54	495.38	315.55	7	<0.001***
Latent means test[b]	50	174.54	3.76	2	0.15	49	190.98	11.15	2	<0.004**
Latent means test after freeing 'sat' mean						48	182.58	2.75	1	0.09 ·

Significance codes: 0.001> '***'; 0.01 >'**'; 0.05 >'*'; 0.1 >'·'

[a,b,c]LRT with respect to a model of (partial) invariance of intercepts

4.7 General discussion

In the present research, we explored the effects of the language of a survey questionnaire on the answers of bilingual respondents.Except for translation issues, the study of how the language of a questionnaire can affect respondents' answers has received little attention in comparative survey methodology. As cross-national comparative survey research expands to populations of study that are culturally diverse, measurement instruments are translated into more languages and more sampled individuals are themselves bilingual. This motivated the study of the potential effects that the language of the survey has on bilingual individuals. A limitation of this study is that the sample size is not large enough to divide the analysis by linguistic group in the bilingual sample. Further research needs to be done on specific cultural groups.

Three specific research objectives were addressed in the present research. The first was whether language effects would emerge in bilingual individuals of cultural backgrounds different from those tested in the majority of published articles (Asian and Hispanic descendants). In our study, participants were Dutch bilingual individuals. The second question was whether language effects would emerge in political constructs, as cultural and self-identity constructs have been more often explored in previous research. The third was to challenge the classical approach of testing for language effects comparing observed means of composite scores by testing whether the correlation of a latent variable in two languages is one.

The procedure we followed was the first step, to test for within-subject measurement equivalence to confirm that our measures in two languages were metric and scalar invariant. Testing for measurement equivalence between languages was seldom performed in past research and it is a prerequisite for statistical comparison of survey items across cultures, languages and groups (Vandenberg & Lance, 2000). The second step was to test for differences in latent correlations and means.

The first conclusion is that the measures we used for the concepts in Model 1: 'Trust and need of change in institutions' and in Model 2:‘Satisfaction and need forchange in politics and the economy’are statistically equivalent across languages.

The second conclusion is thatthe language in a survey questionnaire affectsto some extent the answers of bilingual respondents to political dimensions. We found, in both models, that the correlation between a latent variable measured by the same questions in Dutch and in a different language was not equal to one, but significantly lower. However, factor mean differences did not emerge[15].

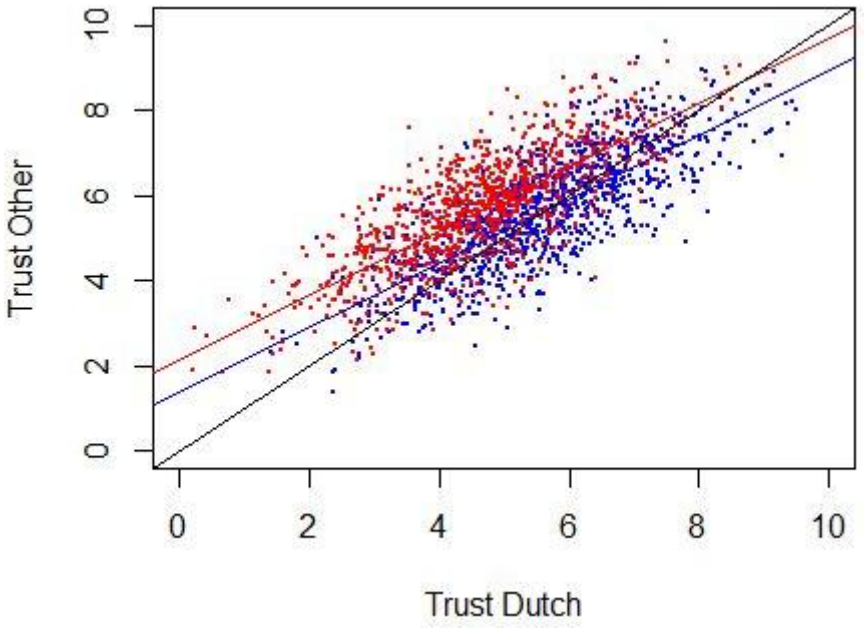
This result indicates that language effects can be present even in cases where significant differences in latent means do not emerge. Latent mean differences indicate a difference in the location

[15] Very small significant latent means were found in Model 2, but they were below the set threshold to consider them relevant.

parameter of the latent variable distribution. This result is illustrated in Figure 15 where we generated random draws from a multivariate normal distribution with the values obtained from the estimation of Model 1 (Appendix 4.2 reports the parameters used in the simulation). In that model, the correlation of the latent variable 'Trust' in Dutch and in a second language was 0.78 and the point estimates of the latent meanswere 5.65 in Dutch and the second language.

The blue cloud of points represents individual scores when the latent means are not significantly different across languages but the correlation is 0.78. The blue line shows the regression line that best fits that cloud of points. The red cloud of points represents the individual scores in a scenario in which the latent meandifferences are significant (4.65) and the correlation is 0.78. Consequently, the red line shows the regression line that best fits that data. Neither the red line nor the blue line overlap with the black diagonal line.The later represents the line that would fit the data when the two variables are the same.

Figure 15. Correlation of two latent variables with and without equal means



By borrowing the theoretical framework of *cultural frame switching* (CFS) (Hong et al., 2000) from cultural psychology, we interpret our results arguing that respondents made use of different conceptual associations in each language. As each language is associated to language specific cultural orientations, our results indicate that respondents shifted their cultural frame of reference to answer in each language.

- Implications for survey methodology

Survey questions are measurement instruments of opinions. If the correlation between thesame latent variable in two languages is not one, apparently it would follow that bilingual individuals have two opinions. The first implication for the design of surveys with

multilingual samples is the decision regarding which language should be given to bilingual respondents. The first possibility would be to give respondents two questionnaires, as we did in this study, and average their opinion. From an operational point of view, this solution is not very optimal e.g. increases costs, increases cognitive burden in the respondent, increases the length of the interview, and introducespotential memory effects, etcetera. A second option (suggested in Richard & Toffoli, 2009) would be to randomize the questionnaires across languages, in a survey like the one presented in this study. This would have meant that a random group of respondents would have answered in Dutch and some othersin a second language. Although this option is statistically sound as differences across languages would cancel out, it is not operational in a comparative survey. The linguistic characteristics of the target population and of the individuals in the sampling frame are in general unknown before data collection. Thus, the size of the random groups would be unknown as well. Moreover, bilingualism implies the dual ability to write, speak, read and listen intwo languages. It also implied the usage of both languages in their main activities and with friends and relatives (Grosjean, 2014). It does not imply that respondents should feel fully comfortable answering certain topics in one or the other language.

We argue that for survey items measuring political dimensions, researchers should choose the language of administration of the questionnaire. The argument is that this fixes the conceptual

associations linked to that language, making them the same as for monolingual respondents.

Appendix 4.1 Development of questions

Three step procedure (Saris & Gallhofer, 2014) to design the survey questions used in the resent research.

Model 1: Institutions: trust and change

Concept-by-postulation 1: Trust in institutions

Concepts-by-intuition:

1) Trust in the parliament.
2) Trust in the police.
3) Trust in the political parties.

Assertions:

1) Respondent trusts the parliament.
2) Respondent trusts the police.
3) Respondent trusts the political parties.

Survey questions:

We will ask some questions about your level of trust in some institutions, 0 indicates complete distrust and 10 complete trust

1) Overall, how much you trust the parliament?

Complete distrust					Neither distrust nor trust					Completely
0	1	2	3	4	5	6	7	8	9	10

2) How much you personally distrust or trust the police?

Complete distrust					Neither distrust nor trust					Completely
0	1	2	3	4	5	6	7	8	9	10

3) How much you personally trust the political parties?

Complete distrust					Neither distrust nor trust					Completely
0	1	2	3	4	5	6	7	8	9	10

Concept-by-postulation 2: Need of change in the institutions

Concepts-by-intuition:

1) Need of change in the parliament.
2) Need of change in the police.
3) Need of change in the political parties.

Assertions:

1) The Dutch parliament needs to change the way it works.
2) The police needs to change the way it works to protect people like you.
3) Political parties need to change the way they work

Survey questions:

The next questions are about change in institutions, 0 indicates you think it does not need to change at all the way it works and 10 indicates it needs to change completely.

1) How much you think that the Dutch parliament needs to change the way it works?

No need to change at all										Completely
0	1	2	3	4	5	6	7	8	9	10

2) How much you think that the police needs to change the way it works to protect people like you?

No need to change at all										Completely
0	1	2	3	4	5	6	7	8	9	10

3) To what extent do political parties need to change the way they work?

No need to change at all										Completely
0	1	2	3	4	5	6	7	8	9	10

Model 2: Politics and the economy: satisfaction and change

Concept-by-postulation 1: Satisfaction with politics and the economy

Concepts-by-intuition:

1) Satisfaction with the economy in the Netherlands
2) Satisfaction with the Dutch government
3) Satisfaction with democracy in the Netherlands

Assertions:

1) Respondent is satisfied with the present state of the economy in the Netherlands.
2) Respondent is satisfied with the way the Dutch government is doing its job
3) Respondent is satisfied with the way democracy works in the Netherlands

Survey questions:

Now we will ask you some questions about your satisfaction with some aspects of politics and the economy. Use a scale from 0 to 10 where 0 means you are completely dissatisfied and 10 means you are completely satisfied.

1) How satisfied are you with the present state of the economy in the Netherlands?

Completely dissatisfied									Completely satisfied	
0	1	2	3	4	5	6	7	8	9	10

2) Overall, how satisfied are you with the way the Dutch government is doing its job?

Completely dissatisfied									Completely satisfied	
0	1	2	3	4	5	6	7	8	9	10

3) And overall, how satisfied are you with the way democracy works in the Netherlands?

Completely dissatisfied									Completely satisfied	
0	1	2	3	4	5	6	7	8	9	10

Concept-by-postulation 2: Need of change in politics and the economy

Concepts-by-intuition:

1) The economic system needs to change in the Netherlands
2) Change in the Dutch government
3) Change in the way democracy works in the Netherlands

Assertions:

1) Respondent is satisfied with the present state of the economy in the Netherlands.
2) The Dutch government needs to change the way it is doing its job.
3) The way democracy works in The Netherlands needs to change.

Survey questions:

We will ask you about the level of change you think some aspects of in politics and the economy need, 0 indicates ‘there is no need at all to change’ and 10 is that ‘it needs to change completely’.

1) To what extent the economic system needs to change?

Not need at all to change										Completely
0	1	2	3	4	5	6	7	8	9	10

2) Overall, to what extent does the Dutch government need to change the way it is doing its job?

Not need at all to change										Completely
0	1	2	3	4	5	6	7	8	9	10

3) To what extent does the way democracy works in The Netherlands needs to change?

Not need at all to change										Completely
0	1	2	3	4	5	6	7	8	9	10

Appendix 4.2 Simulation parameters

We drew 1,000 values from a multivariate normal distribution

$$N_4\left(\mu = \begin{bmatrix}\kappa_1 = 5.65\\ \kappa_3 = 5.65\\ \kappa_2 = 5.83\\ \kappa_4 = 5.83\end{bmatrix}, \Sigma = \begin{bmatrix}1.81\ 1.37 - 1.31 - 1.11\\ 1.37\ 1.73\ - 1.14 - 1.02\\ -1.31 - 1.14\ 2.22\ 1.35\\ -1.11, -1.02, 1.35, 1.20\end{bmatrix}\right)$$

in the case of equal latent means and

$$N_4\left(\mu = \begin{bmatrix}\kappa_1 = 4.65\\ \kappa_3 = 5.65\\ \kappa_2 = 5.83\\ \kappa_4 = 5.83\end{bmatrix}, \Sigma\right)$$

in the case of different latent means using MASS package in R (Venables & Ripley, 2002)

Appendix 4.3. Global fit indices of the models

Model 1. Trust and need of change in institutions

Baseline model: DF = 42; $\chi^2 = 209.9$ ($p = 0.00$); RMSEA = 0.125, 90 % CI for RMSEA = (0.109 ; 0.142); CFI = 0.917, SRMR=0.071.

Baseline model + correlated errors: DF = 40; $\chi^2 = 158.9$ ($p = 0.00$); RMSEA = 0. 108, 90 % CI for RMSEA = (0. 091 ; 0.126); CFI = 0.941, SRMR=0.060.

Factor loadings invariance: DF = 44; $\chi^2 = 165$ ($p = 0.00$); RMSEA = 0.104, 90 % CI for RMSEA = (0. 087 ; 0.121); CFI = 0.940, SRMR=0.63.

Invariance of intercepts: DF = 48; $\chi^2 = 170.8$ ($p = 0.00$); RMSEA = 0.100, 90 % CI for RMSEA = (0.084 ; 0.117); CFI = 0.939, SRMR=0.064.

Test of latent means differences: DF = 50; $\chi^2 = 174.5$ ($p = 0.00$); RMSEA = 0.099, 90 % CI for RMSEA = (0. 083 ; 0.115); CFI = 0.938, SRMR=0.064.

Test of latent correlations = 1: DF = 54; $\chi^2 = 417.3$ ($p = 0.00$); RMSEA = 0.162, 90 % CI for RMSEA = (0.148 ; 0.177); CFI = 0.820, SRMR=0.119.

Model 2. Satisfaction and need of change in politics and the economy

Baseline model: DF = 42; $\chi^2 = 232.8\,(p = 0.00)$; RMSEA = 0.113, 90 % CI for RMSEA = (0.117 ; 0.150); CFI = 0.916, SRMR=0.072.

Baseline model + correlated errors: DF = 39; $\chi^2 = 172.4\,(p = 0.00)$; RMSEA = 0. 116, 90 % CI for RMSEA = (0. 098 ; 0.134); CFI = 0.941, SRMR=0.070.

Factor loadings invariance: DF = 43; $\chi^2 = 175.4\,(p = 0.00)$; RMSEA = 0.110, 90 % CI for RMSEA = (0. 093 ; 0.127); CFI = 0.942, SRMR=0.72.

Invariance of intercepts: DF = 47; $\chi^2 = 179.8\,(p = 0.00)$; RMSEA = 0.105, 90 % CI for RMSEA = (0.089 ; 0.122); CFI = 0.942, SRMR=0.073.

Test of latent means differences: DF = 49; $\chi^2 = 191\,(p = 0.00)$; RMSEA = 0.107, 90 % CI for RMSEA = (0. 091 ; 0.123); CFI = 0.938, SRMR=0.075.

Latent means test after freeing 'sat' mean: DF = 48; $\chi^2 = 182.6\,(p = 0.00)$; RMSEA = 0. 105, 90 % CI for RMSEA = (0.089 ; 0.121); CFI = 0.941, SRMR=0.073.

Test of latent correlations = 1: DF = 54; $\chi^2 = 495.4\,(p = 0.00)$; RMSEA = 0.179, 90 % CI for RMSEA = (0.165 ; 0.194); CFI = 0.806, SRMR=0.239.

Chapter 5

Conclusions

5. Conclusions

The main objective of this dissertation is to contribute to the answer of the research question: How does language in a comparative survey affect equivalence? I studied linguistic equivalence in survey research from three perspectives: survey translation, linguistically diverse countries, and bilingualism.

This chapter provides general conclusions for the dissertation based on the conclusions derived from each of the chapters. For each article, Table 17 provides a summary of the current practices, methodological gap, contribution of the article and conclusions related to the understanding of equivalence with respect to language. The remainder of this chapter explains the summary of Table 17 in detail, extending to the areas where future research is needed. Finally, the chapter closes with a section that intends to provide survey methodologists and practitioners with advice on how to improve the design of multilingual surveys based on the dissertation's findings.

5.1 Survey translation

Survey translation has developed best practice procedures to translate functionally equivalent survey questionnaires. However, in practice, it is a complex challenge to empirically check that the requirements set by the state-of-the-art translation guidelines are fulfilled.

The complexity is that elements which matter to design a good question should be monitored in all participating languages. This is of course a very difficult task without a clear inventory of the elements that should remain fixed when translating survey items. After reviewing best practice procedures to translate and assessing translations of survey items, the conclusion is that a framework that allows us to compare elements of translated questions in a systematic way does not yet exist. Translation assessment methods reviewed in Chapter 2 provide useful information on different translation elements but they focus on some only of the characteristics or content-related features. Assessment methods rely mostly on judgements that may make findings partial or subjective because they depend on the evaluators' knowledge of the survey context or even stylistic preferences about the language.

Table 17. Summary of the thesis contributions and conclusions.

	Current practices, methodological gap, contribution of the thesis and conclusions
Article 1	***Current practice/ evidence***
	A good translation should maintain the same concepts across languages, preserve the item structure and maintain the intended psychometric properties (Janet A. Harkness, Villar, et al., 2010; Janet A. Harkness, 2003)
	Methodological gap
	Guidelines in survey translation do not link assessment criteria and measurement equivalence testing. Empirical assessment of translations is made once data has already been collected
Survey translation	***Contribution of the thesis***
	Propose a systematic procedure to compare versions of a question in different languages before fieldwork. The procedure compares survey feature codes that are predictors of measurement quality, linking translation assessment with measurement equivalence.
	Conclusions
	There are differences across languages that can be prevented using the procedure but there are differences that are unavoidable due to the language's structure. A systematic comparison of translated item properties strengthens equivalence because it provides a scheme of those elements that should remain fixed across languages.
Article 2	***Current practice/ evidence***
	Test for invariance at the country level (Steinmetz, 2011)
	Methodological gap
	In countries where survey instruments are translated into more than one language, it is not tested, but assumed, whether linguistic groups are invariant.
Linguistic groups within countries	***Contribution of the thesis***
	Test for measurement invariance of survey questions taking into account linguistic diversity within countries. Test for invariance, distinguishing the response and cognitive processes to a survey question.
	Conclusions
	Invariance was established (in general) at the cognitive level, whereas differences across groups emerged in their reaction to the measurement method.

Table 17. Cont.

	Current practices, methodological gap, contribution of the thesis and conclusions
Article 3	***Current practice/ evidence***
Bilingual respondents	The language in an interview may activate cultural orientations driving bilingual individuals' responses. A common practice is letting bilingual respondents choose the language of the questionnaire (ESS 2014; OECD 2012). It is also suggested that the language could be randomized (Richard & Toffoli, 2009) Cultural topics are susceptible to language effects (S. X. Chen & Bond, 2010). ***Methodological gap*** It has not been explored whether topics that are not cultural but affected by culture, e.g. political dimensions, are affected by language. ***Contribution of the thesis*** Extent the study of language effects to questions about political dimensions. Study measurement equivalence within an individual in two languages. Once invariance is established, test for language effects using latent variables (correlations in addition to mean differences). ***Conclusions*** Measurement instruments in two languages for the same individual can be equivalent. However, the correlation of a concept is not necessarily equal to one. This implies that the answers can be compared, but that latent opinions, and the associations related to them, are not necessarily the same. For political topics, survey researchers should assign the language of the survey to bilingual individuals. This implicitly homogenizes the associations used by both monolingual and bilingual respondents

In this dissertation, I suggested a procedure for detecting deviations relevant for comparability of different language versions of a survey instrument before it is administered to respondents. The procedure uses a survey featuresinventory that determines the measurement quality (reliability and validity) of the survey items. As measurement quality is estimated with a model that also tests for measurement invariance, the comparability of questions is enhanced when the questions' characteristics remain the same across

languages. With a coding scheme, question similarity is accomplished to the extend the codes remain the same across languages because it implies that regardless of the syntaxes and grammar of the languages, the item features remain the same.

The software Survey Quality Predictor (SQP) (Saris et al., 2011) contains such a coding scheme of item features that predicts the measurement quality of a survey item. The coding process for comparing features of translated items was implemented for a sample of questions in Rounds 5, 6 and 7 of the European Social Survey.

By defining the intended measurement properties of the survey item that should remain constant across languages, the procedure suggested in Chapter 2 of this dissertation was successful at preventing a large number of differences across languages that were not warranted and has helped to better communicate the objectives of survey translation. In the ESS, this led not only to changes in the translation of some items in some languages, but also to better annotations of the source questionnaire and to fewer idiomatic expressions.

In conclusion, multilingual survey questionnaires require researchers to define a priori the elements of the items that should remain fixed after translation processes is not only related to content but also to the formal properties of the survey item. With such an inventory, it is possible to develop procedures for checking that

translation teams align with the requirements set by translation guidelines before the data is collected.

- Future areas of research

Future research should strengthen the link between functional equivalence and translation assessment. Translation assessment procedures need to be developed keeping in mind a framework for statistical equivalence. In this dissertation, I used the framework of equivalence defined by a measurement model. Then, I used an inventory of measurement characteristics that are known to affect estimates of measurement quality (Saris & Gallhofer, 2014) when comparing source and target instruments. Future research would develop from the current inventory of characteristics that are compared, possibly incorporating elements that are related to semantic comparison.

A second line of research is related to managing and analysing large amounts of information derived from survey translation procedures. Large scale cross-national survey projects involve many different languages and some translation problems and solutions could be generalized into families of languages. Some other problems can be very specific for a language. For instance, in Round 5 we learnt from the Lithuanian language that it is not possible to leave alone short texts in the scales such as adverbs without a complement, a grammatical person or a personal pronoun. This result was consistent for several Slavic languages. The evidence gathered by

Round 7 allowed concluding that this was a general issue in Slavic languages. The process has helped us to be aware of those differences and to aid translators in taking better decisions when the form of the question should be adapted. However, work needs to be done to systematize such findings.

A limitation of SQP Coding is inherent to coding procedures: coding can be tiresome and coders should be trained carefully to minimize coding errors. Survey translation could explore tools for computational linguistics technology in which the information about the item characteristics that need to be compared can be extracted automatically.

A third related limitation is in the scope of the approach.As the comparison of codes focuses on the form of the items, content is dismissed. This issue has been tackled by asking national teams during the reporting step how specific formulations were solved. In the case that multilingual corpora for the translation of questionnaires were available, solutions across languages could be better documented.

5.2 Linguistic groups within countries

In order to compare groups using survey data, it is necessary to test for measurement invariance. Current procedures in this respect face two challenges. Firstly, measurement invariance tests using large scale comparative surveys are often limited to testing across

countries. The assumption is that they are homogeneous cultural entities. However, the presence of linguistic groups in multilingual countries makes up culturally diverse countries. In the co-authored article presented in Chapter 3 of this dissertation, we challenge the assumption that cultural groups within a country are invariant and we extend the test for measurement invariance across linguistic groups.

Secondly, an invariance test in multilingual samples should establish whether linguistic groups interpret concepts in the same way, taking into account how the measurement instrument affects the observed responses, that is, a distinction between cognitive and response process in a measurement model should be made.

This distinction cannot be made using the well-established procedure to test for measurement invariance presented in the introductory chapter (Meredith, 1993; Steenkamp & Baumgartner, 1998). It has a known flaw: when the null hypothesis of invariance is rejected, the standard model does not provide information about whether (linguistic) groups are using different interpretations of concepts, if there are differences in the response process, or both. This makes the standard procedure very strict because if differences across groups are only present in the response process, comparisons across groups can be done by correcting differences at this level. Invariance at the cognitive level becomes the only necessary and sufficient condition for cross-cultural comparison of survey data (Saris & Gallhofer, 2014).

In the article, we used an alternative parameterization suggested by Saris and Gallhofer (2014) in which the cognitive and the response processes are distinguished to avoid differences in the reaction to the measurement method confounding differences in the equivalence of the understanding of the concepts across groups. Results show that, in general, linguistic groups exhibited invariance in the cognitive level and they share the same understanding of the concepts asked. In the four studies included in this paper, response differences in the reaction to the method of the questionnaire across groups were more common.

These results derive two main conclusions. Firstly, it cannot be assumed that data coming from linguistically diverse countries can be aggregated without testing for measurement invariance. Secondly, linguistic groups can react in different ways to the survey item. That reaction is not necessarily connected to the way they interpret the concepts, but to the way questions are formulated and presented to the respondent.

- Future areas of research

The distinction between the response part and the cognitive part of a measurement process requires information about the measurement error of the questions. In our paper, we identified the model by having two questions in which the item stem is the same but variations in the measurement method were introduced e.g. the response scale. This implies that each respondent answered two

questions about a same topic. This design is very difficult to be implemented in a socio-political survey e.g. increases costs, increases cognitive burden in the respondent, increases the length of the interview, and introduces potential memory effects. Therefore, instead of using a multiple-items approach, a possibility is to introduce information about measurement error in the model from an external source. This would require the estimation of such a measurement quality.

5.3 Bilingualism

In Chapter 4 of the present research, we explored the effects of the language of a survey questionnaire on the answers of bilingual respondents. Three specific research objectives motivated this study. The first was to investigate if language effects would emerge in answers or opinions about politics, as cultural and self-identity constructs have been more often explored in previous research (S. X. Chen & Bond, 2010; Luna et al., 2008; Ramírez-Esparza et al., 2006; Schwartz et al., 2014). The second motivation was whether language effects would emerge in bilingual individuals from cultural backgrounds different from those tested in the majority of published articles (Asian and Hispanic descendants). In our study, participants were Dutch bilingual individuals.

The second objective was to challenge the classical approach of testing for language effects. Two methodological weaknesses were identified in past research. The first is that language effects have

been assessed without previously establishing statistical equivalence in the measurement instruments, i.e. invariance has been assumed rather than tested. The second is that, in previous research, composite scores' mean differences have been tested.When they were not significant, the conclusion has been that language effects were negligible. In Chapter 4, I tested whether the correlation of a latent variable affecting measures in one language and the other was equal to one and whether the differences in latent means, rather than observed means, were significant. Unless there are composite scores, latent variables are correctedfor measurement error.

Results show that the measures about trust in institutions and political satisfaction are statistically equivalent across languages. However, the language of the questionnaire affected to some extent the opinions of bilingual respondents to political dimensions. The correlation between a latent variable measured by the same questions in Dutch and in a different language was not equal to one, but significantly lower. In line with the theoretical framework of *cultural frame switching* (CFS) developed by cultural psychologists (Hong et al., 2000), I argue that respondents made use of different conceptual associations in each language. As each language is associated to language-specific cultural orientations, results indicate that respondents shifted their cultural frame of reference to answer in each language.

Language effects can be present even in cases when significant differences in latent means do not emerge. Latent mean differences

indicate a difference in the location parameter of the latent variable, however it is possible to arrive at the same mean score in the latent variable from two different conceptual associations.

If the correlation between the same latent variable in two languages is not one, the result is that bilingual individuals have two opinions. The first implication for the definition of fieldwork procedures is to decide which language should be given to bilingual respondents. The resulting conclusion is that for political topics, survey designers should

choose the language of the survey to implicitly fix the cultural framework used by both bilingual and monolingual respondents.

- Future areas of research

The study implemented in this dissertation has a limitation in the sample size of the subgroups. It was not large enough to divide the analysis by linguistic group in the bilingual sample. Further research needs to be done to broaden the analysis of language effects to more cultural groups. Research in cultural psychology has been conducted in cultures --and associated languages-- that are very different to each other such as Chinese and English. Research may also be conducted for languages that are similar to each other, for instance, Catalan and Castilian Spanish in Spain or Ukrainian and Russian in Ukraine. Another possibility is to study cultural groups that are similar, even if the languages come from different

families, for instance, Flanders and French in Belgium or German and French in Switzerland.

A second area of further research is regarding the topics for which language effects are explored. In sociocultural psychology, research has been focused on self-identity, personality and cultural constructs. In this dissertation, I studied two political concepts largely used in Political Science: trust in institutions and satisfaction with politics. Other topics that have been explored in the literature are about self-reported health measures (Elliott et al., 2012; Peytcheva, 2008). Therefore, a systematic study of language effects should cover a larger range of topics asked in attitudinal surveys.

Finally, a third area of further research should explore how the mechanisms of cultural frame switching, CFS, (Hong et al., 2000) work in the context of an interview. There is evidence suggesting that non-verbal cues moderate CFS, such as the interviewers' ethnicity and their accent in the languages (S. X. Chen & Bond, 2010), but little is known about how these mechanisms work in an interview out of a laboratory setting.

5.4 Concluding remarks

Except for translation issues, the study of language effects in survey methodology is relatively new. However, as comparative surveys spread in the social sciences, it is gaining importance. I close this concluding chapter summarizing how the findings of the three

articles compiled in this dissertation help to improve the know-how of crafting multilingual surveys. This is done in Table 18 below.

Table 18. Recommendations for survey methodology

Area of decision	Know-how
Survey translation	Once translations are ready, conduct a pilot study to test for measurement equivalence. If that is not possible: Assess translated questions with a coding system of item characteristics that are known to affect equivalence Use this scheme to communicate to translation teams the objective of the translation and to structure the elements that should remain fixed across languages. When possible, organize issues in families of languages to look for solutions in a transversal way.
Analysis of survey data in multilingual countries	Before aggregating data at the country level, test for measurement invariance across linguistic groups in a country. If invariance is established, go on with testing at the cross-national level. If invariance within a country is rejected, try to incorporate information about the response process in the model. If non-invariance is due to respondents' differential reaction to the method of the survey, this should be corrected. Having done this, it is possible to go on with the cross-national test. If non-invariance is due to a different interpretation of the concepts, exclude that linguistic group from the cross-national test. If possible, go back to the questionnaire design stage to improve the measures.
Bilingualism in a survey	Revise current practices in multilingual countries with respect to the choice of the language in the survey with the objective of standardizing procedures. Train interviewers with respect to language choice. Bilingual respondents should be given the language chosen at the design stage.

Bibliography

Allum, N., Read, S., & Sturgis, P. (2011). Evaluating change in social and political trust in Europe. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 35–53). New York: Routledge Academic.

Alwin, D. F. (2007). *Margins of Error: A Study of Reliability in Survey Measurement*. Hoboken, NJ: John Wiley & Sons.

Andrews, F. M. (1984). Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach. *Public Opinion Quarterly*, *48*(2), 409–442. http://doi.org/10.1086/268840

Asparouhov, T., & Muthén, B. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. http://doi.org/10.1080/10705511.2014.919210

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*(2), 287–311.

Behr, D. (2009). *Translationswissenschaft und international vergleichende Umfrageforschung*. *Leibniz-Institut für Sozialwissenschaften*.

Benet-Martínez, V., & Haritatos, J. (2005). Bicultural Identity Integration (BII): Components and Psychosocial Antecedents. *Journal of Personality*, *73*(4), 1015–1050. http://doi.org/10.1111/j.1467-6494.2005.00337.x

Benet-Martínez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, *75*(3), 729–750.

Benet-Martínez, V., Lee, F., & Leu, J. (2006). Biculturalism and Cognitive Complexity: Expertise in Cultural Representations. *Journal of Cross-Cultural Psychology*, *37*(4), 386–407. http://doi.org/10.1177/0022022106288476

Benet-Martinez, V., Leu, J., Lee, F., & Morris, M. W. (2002). Negotiating Biculturalism: Cultural Frame Switching in Biculturals with Oppositional Versus Compatible Cultural Identities. *Journal of Cross-Cultural Psychology*, *33*(5), 492–516. http://doi.org/10.1177/0022022102033005005

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (2011). *Measurement errors in surveys* (Vol. 173).

John Wiley & Sons.
Blais, A., & Gidengil, E. (1993). Things are not always what they seem: French-English differences and the problem of measurement equivalence. *Canadian Journal of Political Science*, *26*(03), 541–555.
Blalock, H. M. (1990). Auxiliary measurement theories revisited. In J. J. Hox & J. De Jong-Gierveld (Eds.), *Operationalization and Research Strategy*. (pp. 33–49). Amsterdam: Swets & Zeitlinger.
Bollen, K. K. A. (1989). *Structural Equations with Latent Variables*. Wiley-Interscience; 1 edition.
Bond, M. H. (1985). Language as a Carrier of Ethnic Stereotypes in Hong Kong. *The Journal of Social Psychology*, *125*(1), 53–62. http://doi.org/10.1080/00224545.1985.9713508
Bond, M. H., & Yang, K.-S. (1982). Ethnic Affirmation Versus Cross-Cultural Accommodation: The Variable Impact of Questionnaire Language on Chinese Bilinguals from Hong Kong. *Journal of Cross-Cultural Psychology*, *13*(2), 169–185. http://doi.org/10.1177/0022002182013002003
Botha, E. (1968). Verbally Expressed Values of Bilinguals. *The Journal of Social Psychology*, *75*(2), 159–164. http://doi.org/10.1080/00224545.1968.9712488
Botha, E. (1970). The effect of language on values expressed by bilinguals. *Journal of Social Psychology*, *80*(2), 143. Retrieved from http://search.proquest.com/docview/1290717919?accountid=14708
Brislin, R. W. (1970). Back-Translation for Cross-Cultural Research. *Journal of Cross-Cultural Psychology*, *1*(3), 185–216. http://doi.org/10.1177/135910457000100301
Brislin, R. W. (1976). Comparative research methodology: Cross-cultural studies. *International Journal of Psychology*, *11*(3), 215–229. http://doi.org/10.1080/00207597608247359
Brons, C. (2014). *Political discontent in the Netherlands in the first decade of the 21th century*. Tilburg University. Retrieved from https://pure.uvt.nl/portal/files/4230441/Brons_Political_10_10_2014.pdf
Butts, M. M., Vandenberg, R. J., & Williams, L. J. (2006). Investigating the Susceptibility of Measurement Invariance Tests: the Effects of Common Method Variance. *Academy of Management Proceedings*, *2006*(1), D1–D6.

http://doi.org/10.5465/AMBPP.2006.27182126
Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456.
Byrne, B. M., & Van De Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*(2), 107–132.
Byrne, B. M., & Watkins, D. (2003). The Issue Of Measurement Invariance Revisited. *Journal of Cross-Cultural Psychology*, *34*(2), 155–175. http://doi.org/10.1177/0022022102250225
Candell, G. L., & Hulin, C. L. (1986). Cross-Language and Cross-Cultural Comparisons in Scale Translations: Independent Sources of Information about Item Nonequivalence. *Journal of Cross-Cultural Psychology*, *17*(4), 417–440. http://doi.org/10.1177/0022002186017004003
CentERData. (2010). LISS Panel. Retrieved from https://www.lissdata.nl/lissdata/about-panel
Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 464–504.
Chen, S. X., Benet-Martínez, V., & Ng, J. C. K. (2014). Does Language Affect Personality Perception? A Functional Approach to Testing the Whorfian Hypothesis. *Journal of Personality*, *82*(2), 130–143. http://doi.org/10.1111/jopy.12040
Chen, S. X., & Bond, M. H. (2010). Two languages, two personalities? Examining language effects on the expression of personality in a bilingual context. *Personality and Social Psychology Bulletin*, *36*(11), 1514–1528.
Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. http://doi.org/10.1207/S15328007SEM0902_5
Cieciuch, J., Davidov, E., Oberski, D., & Algesheimer, R. (2015). Testing for measurement invariance by detecting local misspecification and an illustration across online and paper-and-pencil samples. *European Political Science*, *14*(4), 521–538. http://doi.org/10.1057/eps.2015.64
Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., &

Schwartz, S. H. (2014). Comparing results of an exact versus an approximate (Bayesian) measurement invariance test: a cross-country illustration with a new scale to measure 19 human values. *Frontiers in Psychology, Forthcoming*.

Cohen, A. B. (2009). Many forms of culture. *American Psychologist*, *64*(3), 194.

Conrad, F. G., & Blair, J. (2004). Data Quality in Cognitive Interviews: The Case of Verbal Reports. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 67–87). Wiley Online Library.

Davidov, E., & De Beuckelaer, A. (2010). How Harmful are Survey Translations? A Test with Schwartz's Human Values Instrument. *International Journal of Public Opinion Research*, *22*(4), 485–510. http://doi.org/10.1093/ijpor/edq030

Davidov, E., Dulmer, H., Schluter, E., Schmidt, P., & Meuleman, B. (2012). Using a Multilevel Structural Equation Modeling Approach to Explain Cross-Cultural Measurement Noninvariance. *Journal of Cross-Cultural Psychology*, *43*(4), 558–575. http://doi.org/10.1177/0022022112438397

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 55–75.

De Leeuw, E. D. (2008). Choosing the method of data collection. In E. D. de Leeuw, J. J. Hox, & D. A. Dillmann (Eds.), *International handbook of survey methodology*. New York, NY: Taylor & Francis Group/Lawrence Erlbaum Associates.

Dean, E., Caspar, R., McAvinchey, G., Reed, L., & Quiroz, R. (2007). Developing a Low-Cost Technique for Parallel Cross-Cultural Instrument Development: The Question Appraisal System (QAS-04). *International Journal of Social Research Methodology*, *10*(3), 227–241. http://doi.org/10.1080/13645570701401032

Dillman, D. A. (2007). *Mail and internet surveys: the tailored design method*. Wiley & Sons.

Dillman, D. A., Smyth, J. D., & Melani, L. (2011). *Internet, mail, and mixed-mode surveys: the tailored design method*. Wiley & Sons Toronto.

Dixon, D. J. (2007). The effects of language priming on independent and interdependent self-construal among Chinese university students currently studying English. *Current*

Research in Social Psychology, *13*, 1–9.
Dorer, B. (2011). *Advance translation in the 5th round of the European Social Survey (ESS)* (FORS Working Paper Series No. 2011-4.). Lausanne.
Dorer, B. (2012). *Round 6 Translation Guidelines.* Mannheim.
Dorer, B. (2013a). *ESS Translation Expert Task Group Meeting*. Mannheim.
Dorer, B. (2013b). *Report on Translation Expert Task Group Meeting. ESS DACE Deliverable*. Mannheim, Germany.
Easton, D. (1975). A Re-Assessment of the Concept of Political Support. *British Journal of Political Science*, *5*(4), 435–457. http://doi.org/10.2307/193437
Elliott, M. N., Edwards, W. S., Klein, D. J., & Heller, A. (2012). Differences by Survey Language and Mode among Chinese Respondents to a CAHPS Health Plan Survey. *Public Opinion Quarterly*, *76*(2), 238–264. http://doi.org/10.1093/poq/nfs020
Ellis, N. (1992). Linguistic relativity revisited: The bilingual word-length effect in working memory during counting, remembering numbers, and mental calculation. *Advances in Psychology*, *83*, 137–155.
Erkut, S., Alarcón, O., Coll, C. G., Tropp, L. R., & García, H. A. V. (1999). The dual-focus approach to creating bilingual measures. *Journal of Cross-Cultural Psychology*, *30*(2), 206–218.
Eurobarometer, S. (2012). *Public opinion in the European Union*.
European Social Survey. (2005). ESS Round 2: European Social Survey Round 2 Data. Bergen: Norwegian Social Science Data Services, Norway – Data Archive and distributor of ESS data.
European Social Survey. (2014). *ESS Round 7 Translation Guidelines*. London: City University London.
European Social Survey. (2015a). ESS Round 7: European Social Survey Round 7 Data. Norwegian Social Science Data Services, Norway – Data Archive and distributor of ESS data.
European Social Survey. (2015b). *Round 8 Survey Specification for ESS ERIC Member, Observer and Guest countries*. London: City University London. Retrieved from http://www.europeansocialsurvey.org/docs/round8/ESS8_project_specification.pdf
European Social Survey. (2016). European Social Survey. Retrieved March 13, 2014, from http://www.europeansocialsurvey.org/

European Values Survey. (2010). *EVS 2008 Guidelines and Recommendations*. Bonn: GESIS – Technical Reports 2010/16. Retrieved from http://www.europeanvaluesstudy.eu/page/data-and-documentation-survey-2008.html

Fitzgerald, R., Widdop, S., Gray, M., & Collins, D. (2011). Identifying Sources of Error in Cross-national Questionnaires: Application of an Error Source Typology to Cognitive Interview Data. *Journal of Official Statistics*, *27*(4), 569–599.

Fleischman, H. L., Hopstock, P. J., Pelczar, M. P., & Shelley, B. E. (2010). *Highlights From PISA 2009: Performance of U.S. 15-Year- Old Students in Reading, Mathematics, and Science Literacy in an International Context (NCES 2011-004). Technical Report*. Washington, DC: U.S. Government Printing Offi.

Goerman, P. L., & Caspar, R. A. (2010). Managing the Cognitive Pretesting of Multilingual Survey Instruments: A Case Study of Pretesting of the U.S. Census Bureau Bilingual Spanish/English Questionnaire. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 75–90). John Wiley & Sons, Inc. http://doi.org/10.1002/9780470609927.ch5

Grosjean, F. (2014). Bicultural bilinguals. *International Journal of Bilingualism* . http://doi.org/10.1177/1367006914526297

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Psychology Press.

Harkness, J. A. (1998). *Cross-cultural survey equivalence*. Mannheim, Germany: ZUMA.

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Hoboken: Wiley & Sons.

Harkness, J. A. (2005). SHARE translation procedures and translation assessment. In A. Borsch-Supan & H. Jurges (Eds.), *The Survey of Health, Aging, and Retirement in Europe - Methodology* (pp. 24–27). Mannheim, Germany: MEA.

Harkness, J. A., Bilgen, I., Córdova Cazar, A. L., Cibelli, K., Huang, L., Miller, D., … Villar, A. (2011). Questionnaire Design. In Survey Research Center (Ed.), *Guidelines for Best Practice in Cross-Cultural Surveys* (3rd ed., p. 725). Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved from

http://www.ccsg.isr.umich.edu/
Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L., Mohler, P. P., … Smith, T. W. (2010). Comparative Survey Methodology. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 1–16). John Wiley & Sons, Inc. http://doi.org/10.1002/9780470609927.ch1
Harkness, J. A., Pennell, B.-E., & Schoua-Glusberg, A. (2004). Survey Questionnaire Translation and Assessment. In *Methods for Testing and Evaluating Survey Questionnaires* (pp. 453–473). John Wiley & Sons, Inc. http://doi.org/10.1002/0471654728.ch22
Harkness, J. A., & Schoua-Glusberg, A. (1998). Questionnaires in translation. *ZUMA-Nachrichten Spezial*, *3*(1), 87–127.
Harkness, J. A., Villar, A., & Edwards, B. (2010). Translation, Adaptation, and Design. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 115–140). John Wiley & Sons, Inc. http://doi.org/10.1002/9780470609927.ch7
Harzing, A.-W. (2005). Does the Use of English-language Questionnaires in Cross-national Research Obscure National Differences? *International Journal of Cross Cultural Management* , *5* (2), 213–224. http://doi.org/10.1177/1470595805054494
Harzing, A.-W. (2006). Response Styles in Cross-national Survey Research: A 26-country Study. *International Journal of Cross Cultural Management*, *6*(2), 243–266. http://doi.org/10.1177/1470595806066332
Hendriks, F. (2009). Contextualizing the Dutch drop in political trust: connecting underlying factors. *International Review of Administrative Sciences* , *75* (3), 473–491. http://doi.org/10.1177/0020852309337686
Hong, Y., Morris, M. W., Chiu, C., & Benet-Martínez, V. (2000). Multicultural minds: A dynamic constructivist approach to culture and cognition. *American Psychologist*, *55*(7), 709–720. http://doi.org/10.1037/0003-066X.55.7.709
Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3-4), 117–44. http://doi.org/10.1080/03610739208253916
Hox, J. J. C. M., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a

Bayesian perspective. *Survey Research Methods; Vol 6, No 2 (2012)*. Retrieved from https://ojs.ub.uni-konstanz.de/srm/article/view/5033

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453.

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. http://doi.org/10.1177/0022002185016002001

Hunt, E., & Agnoli, F. (1991). The Whorfian hypothesis: A cognitive psychology perspective. *Psychological Review*, *98*(3), 377–389. http://doi.org/10.1037/0033-295X.98.3.377

Inglehart, R. (1977). Political Dissatisfaction and Mass Support for Social Change in Advanced Industrial Society. *Comparative Political Studies* , *10* (3), 455–472. http://doi.org/10.1177/001041407701000308

Ji, L., Zhang, Z., & Nisbett, R. E. (2004). Is It Culture or Is It Language? Examination of Language Effects in Cross-Cultural Research on Categorization. *Journal of Personality and Social Psychology*, *87*(1), 57–65. http://doi.org/10.1037/0022-3514.87.1.57

John, O. P., Goldberg, L. R., & Angleitner, A. (1984). Better than the alphabet: Taxonomies of personality-descriptive terms in English, Dutch, and German. *Personality Psychology in Europe*, *1*.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426. http://doi.org/10.1007/BF02291366

Jöreskog, K. G., & Sörbom, D. (2004). LISREL 8.7 for Windows. *Lincolnwood, IL. URL Http://www. Ssicentral. Com/lisrel*, Jöreskog, K. G., & Sörbom, D. (2004). LISREL 8.7 f.

Jöreskog, K. G., & Van Thillo, M. (1973). LISREL. Department of Statistics: University of Uppsala.

Kaase, M., Newton, K., & Scarbrough, E. (1997). Beliefs in Government. *Politics*, *17*(2), 135–139. http://doi.org/10.1111/1467-9256.00044

Kemmelmeier, M., & Cheng, B. Y.-M. (2004). Language and Self-Construal Priming: A Replication and Extension in a Hong Kong Sample. *Journal of Cross-Cultural Psychology*, *35*(6), 705–712. http://doi.org/10.1177/0022022104270112

Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The foreign-language effect thinking in a foreign tongue reduces decision biases. *Psychological Science*, *23*(6), 661–668.
Költringer, R. (1995). Measurement quality in Austrian personal interview surveys. *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments*, 207–224.
Krosnick, J. (1990). The impact of satisficing on survey data quality. In *Proceedings of the Bureau of the Census 1990 Annual Research Conference* (pp. 835–845).
Krosnick, J. A., & Fabrigar, L. R. (1997). Designing Rating Scales for Effective Measurement in Surveys. In L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey Measurement and Process Quality* (pp. 141–164). Hoboken, NJ, USA: John Wiley & Sons, Inc. http://doi.org/10.1002/9781118490013
Lechuga, J. (2008). Is Acculturation a Dynamic Construct?: The Influence of Method of Priming Culture on Acculturation. *Hispanic Journal of Behavioral Sciences*, *30*(3), 324–339. http://doi.org/10.1177/0739986308319570
Lessler, J. T., & Forsyth, B. H. (1996). A coding system for appraising questionnaires.
Levi, M., & Stoker, L. (2000). Political Trust and Trustworthiness. *Annual Review of Political Science*, *3*, 475–507. http://doi.org/10.1146/annurev.polisci.3.1.475
Luna, D., Ringberg, T., & Peracchio, L. A. (2008). One Individual, Two Identities: Frame Switching among Biculturals. *Journal of Consumer Research*, *35*(2), 279–293.
Mallinckrodt, B., & Wang, C.-C. (2004). Quantitative Methods for Verifying Semantic Equivalence of Translated Research Instruments: A Chinese Version of the Experiences in Close Relationships Scale.
Marian, V., & Kaushanskaya, M. (2004). Self-construal and emotion in bicultural bilinguals. *Journal of Memory and Language*, *51*(2), 190–201. http://doi.org/10.1016/j.jml.2004.04.003
Marian, V., & Neisser, U. (2000). Language-Dependent Recall of Autobiographical Memories. *Journal of Experimental Psychology: General*, *129*(3), 361–368. http://doi.org/10.1037/0096-3445.129.3.361
Marsh, H. W., & Byrne, B. M. (1993). Confirmatory Factor

Analysis of Multitrait-Multimethod Self-concept Data: Between-group and Within-group Invariance Constraints. *Multivariate Behavioral Research*, *28*(3), 313–449. http://doi.org/10.1207/s15327906mbr2803_2

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. http://doi.org/10.1007/BF02294825

Meuleman, B. (2012). When are item intercept differences substantively relevant in measurement invariance testing? In S. Salzborn, E. Davidov, & J. Reinecke (Eds.), *Methods, Theories, and Empirical Applications in the Social Sciences* (pp. 97–104). Wiesbaden: Springer VS. http://doi.org/10.1007/978-3-531-18898-0_13

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: how many countries are needed for accurate multilevel SEM? *Survey Research Methods; Vol 3, No 1 (2009)*. Retrieved from https://ojs.ub.uni-konstanz.de/srm/article/view/666

Minkov, M. (2007). *What makes us different and similar: A new interpretation of the World Values Survey and other cross-cultural data*. Klasika i Stil Publishing House.

Mohler, P. P., & Johnson, T. P. (2010). Equivalence, Comparability, and Methodological Progress. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 17–29). John Wiley & Sons, Inc. http://doi.org/10.1002/9780470609927.ch2

Mohler, P. P., Pennell, B.-E., & Hubbard, F. (2008). Survey documentation: Toward professional knowledge management in sample surveys. In E. D. de Leeuw, J. J. Hox, & D. A. Dillmann (Eds.), *International handbook of survey methodology* (pp. 403–420). New York, NY: European Association of Methodology/Lawrence Erlbaum Associates.

Mohler, P. P., & Uher, R. (2003). Documenting comparative surveys for secondary analysis. In J. A. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (Vol. 325, p. 311). Wiley-Interscience.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics*. Chestnut Hill, MA. Retrieved from http://timss.bc.edu/timss2011/international-results-mathematics.html

Muthén, B., & Asparouhov, T. (2013). BSEM measurement

invariance analysis. *Mplus Web Notes*, *17*, 1–48.
Neijens, P. (1987). *Choice questionnaire : design and evaluation of an instrument for collecting informed opinions of a population*. Amsterdam: Free University Press Amsterdam. Retrieved from http://library.wur.nl/WebQuery/clc/242166
Nida, E. A. (1964). *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. Brill Archive.
Northrop, F. S. C. (1947). *The logic of the sciences and the humanities.* New York, NY: Macmillan.
Oberski, D. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*(1), 45–60.
Oberski, D., Saris, W. E., & Hagenaars, J. A. P. (2007). Why are there differences in measurement quality across countries. *Measuring Meaningful Data in Social Research. Acco, Leuven*.
Oberski, D., Saris, W. E., & Hagenaars, J. A. P. (2010). Categorization Errors and Differences in the Quality of Questions in Comparative Surveys. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 435–453). John Wiley & Sons, Inc. http://doi.org/10.1002/9780470609927.ch23
OECD. (2012). *PISA 2012 Technical Report*. Paris. Retrieved from http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf
Oyserman, D., & Lee, S. W. S. (2008). Does culture influence what and how we think? Effects of priming individualism and collectivism. *Psychological Bulletin*, *134*(2), 311.
Pan, Y., & De La Puente, M. (2005). Census Bureau guideline for the translation of data collection instruments and supporting materials: Documentation on how the guideline was developed. *Survey Methodology*, 6.
Pan, Y., Landreth, A., Hinsdale, M., Park, H., & Schoua-Glusberg, A. (2007). Methodology for cognitive testing of translations in multiple languages. In *American Association for Public Opinion Research conference, Anaheim, CA*.
Payne, S. L. (1951). *The Art of Asking Questions*. Princeton, New Jersey: Princeton University Press.
Pennell, B.-E., Harkness, J. A., Levenstein, R., & Quaglia, M. (2010). Challenges in Cross-National Data Collection. In

Survey Methods in Multinational, Multiregional, and Multicultural Contexts (pp. 269–298). John Wiley & Sons, Inc. http://doi.org/10.1002/9780470609927.ch15

Perunovic, E., Wei, Q., Heller, D., & Rafaeli, E. (2007). Within-Person Changes in the Structure of Emotion: The Role of Cultural Identification and Language. *Psychological Science*, *18*(7), 607–613. http://doi.org/10.1111/j.1467-9280.2007.01947.x

Peytcheva, E. (2008). Language of administration as a cause of measurement error. In *AAPOR*. New Orleans.

Pierson, H. D., & Bond, M. H. (1982). How Do Chinese Bilinguals Respond To Variations of Interviewer Language and Ethnicity? *Journal of Language and Social Psychology*, *1*(2), 123–139. http://doi.org/10.1177/0261927X8200100203

PISA. (2010). *Translation and Adaptation Guidelines For PISA 2012*. Budapest: National Project Managers' Meeting. Retrieved from https://www.oecd.org/pisa/pisaproducts/pisa2012translationmanualsandguidelines.htm

Pornprasertmanit, S., Miller, P., Schoemann, A., & Rosseel, Y. (2014). semTools: Useful tools for structural equation modeling. Retrieved from http://cran.r-project.org/package=semTools

R Core Team. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.r-project.org/

Ramírez-Esparza, N., Gosling, S. D., Benet-Martínez, V., Potter, J. P., Pennebaker, J. W., & Ramírez-Esparza, J. W. (2006). Do bilinguals have two personalities? A special case of cultural frame switching. *Journal of Research in Personality*, *40*(2), 99–120.

Reeskens, T., & Hooghe, M. (2007). Cross-cultural measurement equivalence of generalized trust. Evidence from the European Social Survey (2002 and 2004). *Social Indicators Research*, *85*(3), 515–532. http://doi.org/10.1007/s11205-007-9100-z

Richard, M.-O., & Toffoli, R. (2009). Language influence in responses to questionnaires by bilingual respondents: A test of the Whorfian hypothesis. *Impact of Culture on Marketing Strategy*, *62*(10), 987–994. http://doi.org/10.1016/j.jbusres.2008.10.016

Ross, M., Xun, W. Q. E., & Wilson, A. E. (2002). Language and the

Bicultural Self. *Personality and Social Psychology Bulletin*, *28*(8), 1040–1050. http://doi.org/10.1177/01461672022811003
Rosseel, Y. (2012). {lavaan}: An {R} Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/
Sam, D. L., & Berry, J. W. (2010). Acculturation: When Individuals and Groups of Different Cultural Backgrounds Meet. *Perspectives on Psychological Science*, *5*(4), 472–481. http://doi.org/10.1177/1745691610373075
Sanchez, M. E. (1992). Effects of Questionnaire Design on the Quality of Survey Data. *Public Opinion Quarterly*, *56*(2), 206–217. http://doi.org/10.1086/269311
Saris, W. E. (1982a). Different questions, different variables? In C. Fornell (Ed.), *A second generation of multivariate analysis. 2. Measurement and evaluation* (First, Vol. 2). New York: Praeger Publishers.
Saris, W. E. (1982b). Linear structural relations. In C. Fornell (Ed.), *A second generation of multivariate analysis: Methods* (First, Vol. 1). New York: Praeger Publishers.
Saris, W. E. (1988). *Variation in Response Functions: a source of measurement error in survey research*. Amsterdam: Sociometric Research Foundation.
Saris, W. E. (2012). Discussion Evaluation Procedures for Survey Questions. *Journal of Official Statistics*, *28*(4), 537.
Saris, W. E., & Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In P. P. Biemer, R. M. Groves, N. A. Lyberg, L. E. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 575–597). New York: JohnWiley & Sons, Inc.
Saris, W. E., & Andrews, F. M. (2004). Evaluation of Measurement Instruments Using a Structural Modeling Approach. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 575–597). John Wiley & Sons, Inc. http://doi.org/10.1002/9781118150382.ch28
Saris, W. E., & Gallhofer, I. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. *Wiley Series in Survey Methodology* (Vol. 548). John Wiley & Sons. http://doi.org/10.1111/j.1751-5823.2008.00054_20.x
Saris, W. E., & Gallhofer, I. (2014). *Design, Evaluation, and*

Analysis of Questionnaires for Survey Research (Second Edi). John Wiley & Sons.

Saris, W. E., Oberski, D., Revilla, M., Zavala-Rojas, D., Lilleoja, L., Gallhofer, I., & Gruner, T. (2011). *The development of the program SQP 2.0 for the prediction of the quality of survey questions"* (RECSM Working Paper No. 24). Barcelona. Retrieved from http://www.upf.edu/survey/_pdf/RECSM_wp024.pdf

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. http://doi.org/10.1080/10705510903203433

Scherpenzeel, A. C. (1995). *A Question of Quality: evaluating survey questions by multitrait-multimethod studies*. University of Amsterdam.

Scherpenzeel, A. C., & Saris, W. E. (1997). The Validity and Reliability of Survey Questions A Meta-Analysis of MTMM Studies. *Sociological Methods & Research*, *25*(3), 341–383.

Scheuch, E. K. (1993). The cross-cultural use of sample surveys: problems of comparability. *Historical Social Research/Historische Sozialforschung*, 104–138.

Schrauf, R. W., & Rubin, D. C. (2000). Internal languages of retrieval: The bilingual encoding of memories for the personal past. *Memory & Cognition*, *28*(4), 616–623.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Schwartz, S. J., Benet-Martínez, V., Knight, G. P., Unger, J. B., Zamboanga, B. L., Des Rosiers, S. E., … Szapocznik, J. (2014). Effects of language of assessment on the measurement of acculturation: Measurement equivalence and cultural frame switching. *Psychological Assessment*, *26*(1), 100–114. http://doi.org/http://psycnet.apa.org/doi/10.1037/a0034717

Schwartz, S. J., Unger, J. B., Zamboanga, B. L., & Szapocznik, J. (2010). Rethinking the concept of acculturation: Implications for theory and research. *2American Psychologist*, *65*(4), 237–251. http://doi.org/10.1037/a0019330

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology*, *21*(2), 277–287.

Segalowitz, N., Hulstijn, J., Kroll, J. F., & de Groot, A. M. B. (2005). Handbook of bilingualism: Psycholinguistic

approaches. *Handbook of Bilingualism: Psycholinguistic Approaches*.
Smith, T. W. (2004). Developing and Evaluating Cross-National Survey Instruments. In *Methods for Testing and Evaluating Survey Questionnaires* (pp. 431–452). John Wiley & Sons, Inc. http://doi.org/10.1002/0471654728.ch21
Sörbom, D. (1982). Structural equation models with structured means. In K. G. Jöreskog & H. O. Wold (Eds.), *Systems under indirect observation*. Amsterdam.
Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–107.
Steinmetz, H. (2011). Estimation and Comparison of Latent Means Across Cultures. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications.* (pp. 85–116). New York: Routledge Academic.
Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. Jossey-Bass.
Toffoli, R., & Laroche, M. (2002). Cultural and language effects on Chinese bilinguals' and Canadians' responses to advertising. *International Journal of Advertising*, *21*(4), 505–524. http://doi.org/10.1080/02650487.2002.11104948
Tourangeau, R., Rips, L. J., & Kenneth, R. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.
Trafimow, D., Silverman, E. S., Fan, R. M.-T., & Fun Law, J. S. (1997). The Effects of Language and Priming on the Relative Accessibility of the Private Self and the Collective Self. *Journal of Cross-Cultural Psychology*, *28*(1), 107–123. http://doi.org/10.1177/0022022197281007
Triandis, H. C., Davis, E. E., Vassiliou, V., & Nassiakou, M. (1965). *Some Methodological Problems Concerning Research Negotiations Between Monoinguals*.
Tyson, G. A., Doctor, E. A., & Mentis, M. (1988). A Psycholinguistic Perspective on Bilinguals' Discrepant Questionnaire Responses. *Journal of Cross-Cultural Psychology*, *19*(4), 413–426. http://doi.org/10.1177/0022022188194002
Uskul, A. K., Oyserman, D., & Schwarz, N. (2010). Cultural Emphasis on Honor, Modesty, or Self-Enhancement:

Implications for the Survey-Response Process. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 191–201). John Wiley & Sons, Inc. http://doi.org/10.1002/9780470609927.ch11

Van De Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthen, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance . *Frontiers in Psychology* . Retrieved from http://journal.frontiersin.org/article/10.3389/fpsyg.2013.00770

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial "Measurement Invariance." *Frontiers in Psychology*, *6*(1064). http://doi.org/10.3389/fpsyg.2015.01064

Van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research* (Vol. 1). Thousand Oaks, CA, US: Thousand Oaks, CA, US: Sage Publications, Inc.

Van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, *54*(2), 119–135. http://doi.org/10.1016/j.erap.2003.12.004

Van der Veld, W. M. (2006). *The survey response dissected. A new theory about the survey response process.* University of Amsterdam.

Van der Veld, W. M., & Saris, W. E. (2011). Causes of generalized social trust. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 207–247). New York: Routledge Academic.

Van der Veld, W. M., Saris, W. E., & Satorra, A. (2008). Judgement Rule Aid for Structural Equation Models.

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response Styles in Rating Scales: Evidence of Method Bias in Data From Six EU Countries. *Journal of Cross-Cultural Psychology*, *35*(3), 346–360. http://doi.org/10.1177/0022022104264126

Van Meurs, A., & Saris, W. E. (1990). Memory effects in MTMM studies. In W. E. Saris & A. Munnich (Eds.), *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments* (Vol. 1). Budapest: Eötvös University Press.

Vandenberg, R. J. (2002). Toward a Further Understanding of and

Improvement in Measurement Invariance Methods and Procedures. *Organizational Research Methods*, *5*(2), 139–158. http://doi.org/10.1177/1094428102005002001

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. http://doi.org/10.1177/109442810031002

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Villar, A. (2009). *Agreement answer scale design for multilingual surveys: Effects of translation-related changes in verbal labels on response styles and response distributions*. University of Nebraska. Retrieved from http://digitalcommons.unl.edu/sramdiss/3

Watkins, D., & Gerong, A. (1999). Language of Response and the Spontaneous Self-Concept: A Test of the Cultural Accommodation Hypothesis. *Journal of Cross-Cultural Psychology*, *30*(1), 115–121. http://doi.org/10.1177/0022022199030001007

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications.

Willis, G. B., Kudela, M. S., Levin, K., Norberg, A., Stark, D. S., Forsyth, B. H., … Hartman, A. M. (2010). Evaluation of a Multistep Survey Translation Process. In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (pp. 141–156). John Wiley & Sons, Inc. http://doi.org/10.1002/9780470609927.ch8

Yang, K.-S., & Bond, M. H. (1980). Ethnic Affirmation by Chinese Bilinguals. *Journal of Cross-Cultural Psychology*, *11*(4), 411–425. http://doi.org/10.1177/0022022180114002

Yoon, K.-I. (2010). *Political culture of individualism and collectivism*. The University of Michigan.

Zavala-Rojas, D. (2012). *Evaluation of the concepts "Trust in institutions" and "Trust in authorities" (European Social Survey Deliverable 12.4: Evaluation of questions and concepts - report 2.* (RECSM Working Paper 29). Retrieved from http://www.upf.edu/survey/_pdf/RECSM_wp029.pdf

Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E. (2015). The

comparability of the universalism value over time and across countries in the European Social Survey: exact vs. approximate measurement invariance. *Frontiers in Psychology*. Retrieved from http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00733