UNIVERSITAT POMPEU FABRA

RECSM
Research and Expertise Centre for Survey Methodology

RECSM Working Paper Number 8

2009

Universitat Pompeu Fabra - Research and Expertise Centre for Survey Methodology
Edifici de França - Despatxs 70.374-70.382
Passeig de Circumval·lació, 8 - 08003 Barcelona
Tel.: +34 93 542 11 63 web: http://www.upf.edu/survey/ e-mail: recsm@upf.edu

Causes of Generalized Social Trust:
An Innovative Cross-National Evaluation

William M. van der Veld. Radboud Universiteit Nijmegen (Netherlands).
William.vanderVeld@socsci.ru.nl

Willem E. Saris. ESADE, Universitat Ramon Llull and Universitat Pompeu Fabra (Spain)
w.saris@telefonica.es

Author Biographies:

William van der Veld is a specialist in the field of structural equation modeling with a special interest in model testing. He received his PhD, Cum Laude, from the University of Amsterdam. His latest achievement is the development of a software program called JRule, which simplifies model evaluation using an approximate test and taking the power of the test into account. He is a member of the coordinating team of the Dutch Association for Survey Research (NPSO).

Willem Saris is a laureate of the Descartes Prize 2005, for the best collaborative research. In 2009 he received the Helen Dinerman award by the World Association of Public Opinion Research in recognition to his extensive contributions to the fields of public opinion research and development of survey research methods. He is also the president and founder of the European Survey Research Association and a member of the central coordinating team (CCT) of the European Social Survey.

Acknowledgements:

We would like to thank Peter Schmidt, Fons van de Vijver, and William Burk for their valuable suggestions on how to improve this chapter.

Abstract

In this chapter we want to establish whether the 3-item scale to measure generalized social trust (GST) in the European Social Survey (ESS) can be meaningfully used in comparative research. If so, we also want to study the causes of GST cross-nationally; building upon the work by Delhey & Newton (2005). The standard procedure to assess the comparability of a survey measure is by means of measurement equivalence tests, a specific form of multi-group confirmatory factor analyses (MG-CFA). In general it is quite difficult to evaluate these models, i.e. it is not easy to decide when a model fits the data, how and where a model should be improved, and when the improvements are sufficient. In this chapter we introduce two innovations for testing measurement equivalence of survey measures. One innovation involves an alternative procedure to evaluate structural equation models (Saris, Satorra, and Van der Veld, 2009). This procedure is implemented in a software program called JRule which is developed to detect misspecifications in SEM models taking into account the power of the test. The other innovation concerns the separation of error components and the substantive components in the response, which makes it possible to test for measurement equivalence after correction for random and systematic measurement error. Applying these innovations in our study resulted in evidence that the 3-item measure for GST is scalar invariant in all 19 countries that we analyzed.

1. Introduction

What do the following studies have in common: Adam (2008), Herreros & Criado (2008), Kaasa & Parts (2008), Zmerli & Newton (2006), Delhey & Newton (2005), Letki & Evans (2005), Rothstein & Uslaner (2005), Paxton (2002), Inglehart (1999), Kaase (1999), and Knack & Keefer (1997)? In all these studies it is assumed that the measure of generalized social trust (GST) is meaningfully comparable across countries. In addition, it is also assumed that the endogenous and exogenous variables in those studies are measured without error. If any of these assumptions do not hold then conclusions from these studies will be questionable, to say the least. Because measurement without error is very unlikely, one should correct for measurement error. Correction for measurement error can be done in various ways. In this chapter we will discuss how this can be done using estimates from a multitrait multimethod experiment. Correction for measurement error is a necessary step in any comparative study, it does not, however, ensure equivalence of the measures. In order to test whether survey measures are equivalent we have to assess the measurement invariance (Meredith, 1993).

The common procedure to test for invariance of measures is by means of a multi-group confirmatory factor analysis (MG-CFA). Testing in structural equation modeling has become rather difficult to comprehend with the introduction of so many goodness of fit measures (Marsh, Hau, and Wen, 2004). How should we evaluate structural equation models, and what to do if a model is rejected? Where should we begin to improve the model? Usually, there are many possibilities for improvement and each one will have an effect on the set of countries that can be meaningfully compared. Because of the many possibilities, we will suggest an analytic strategy that can guide this process. The strategy requires that JRule (Van der Veld, Saris, and Satorra, 2008) is used to evaluate the MG-CFA model. JRule is primarily developed to detect misspecifications in SEM models. Standard procedures to evaluate model fit are affected by the power of the test and this is not limited to the CHI2 alone, but holds true for other goodness-of-fit measures too (Saris et al., 2009). JRule can detect misspecifications taking into account the power of the test, which is not possible with any other SEM software. To be clear, JRule does not perform a global model evaluation, but it judges whether constrained parameters are constrained to the ‘correct’ values. We will explain this procedure in more detail later in this chapter.

It is not only the standard model evaluation procedure which we suggest to change. The standard model (Meredith, 1993) that is used to assess the cross-national equivalence is flawed too. The reason that the standard model is flawed is because the unique factors in the common factor model are confounded with random measurement error and also that the common (substantive) factor is confounded with systematic measurement error. Therefore the invariance restrictions will lead to wrong conclusions if both random and systematic measurement error components are not the same across countries. We suggest to make a distinction between the unique components in the indicators and the random errors in the indicators, as well as between the systematic error component in the indicators and the common factor. The same was also suggested by Saris & Gallhofer (2007) and by Millsap & Meredith (2007), the latter two authors did however not elaborate on this, nor did they empirically make this distinction.

It is the goal of this chapter to apply these methodological innovations to test the cross-national equivalence of GST in the European Social Survey (ESS) and also to test the causes of GST. Reeskens & Hooghe (2008) have already discussed the cross-national equivalence of generalized trust in the ESS and found that the measure of GST should not be used in comparative studies. We hope to arrive at more rosy conclusions using our improved methodology. Delhey & Newton (2003) have already discussed the causes of social trust and found that at the individual level there were large cross-national differences. We believe that their conclusions are odd. Therefore, we will test whether the same causes are at work in all countries and whether the effect of each cause is more or less the same in all countries. We therefore question previous findings and expect that the causes of GST are the same across all countries and that the effect of each cause is also more or less the same.

The structure of this chapter is as follows, we will first introduce the procedures to test for configural, metric, and scalar invariance. After that we will introduce how we correct for random and systematic measurement error in these invariance tests. Next we will suggest an analytical strategy for the assessment of measurement invariance. This is followed by an introduction to a new procedure to evaluate structural equation models: *the detection of misspecifications using JRule*. Finally, we will apply these methodological improvements and test whether the ESS measure of GST can be meaningfully used in comparative research and if so, whether the causes of GST are the same across the set of countries we analyze.

2. Innovations

2.1. The standard procedure for measurement invariance tests

The standard test has been explained elsewhere in the book, so we skip most of the details. Tests of measurement invariance put restrictions on the measurement model. The basic measurement model is presented below. The superscript is used to indicate different countries, i.e. to indicate that this is a multi-group model.

$$\mathbf{y}^{(n)} = \boldsymbol{\tau}^{(n)} + \boldsymbol{\Lambda}^{(n)}\mathbf{f}^{(n)} + \boldsymbol{\delta}^{(n)} \qquad (1)$$

In this model **y** is the vector of observed variables, **τ** is a vector of intercepts of the observed variables**, f** is a vector of latent variables, **δ** is a vector of disturbance terms of the observed variables, and **Λ** is the matrix of relationships between the observed and latent variables, i.e. the loadings. It is assumed that the mean of all disturbance terms **δ** is zero and that the covariance among the disturbance terms **δ** as well as between the between the disturbance terms and the common factors (**f**) are zero. If these assumptions hold, then the expected value of **y** can be expressed as:

$$\boldsymbol{\mu}_{\mathbf{y}}^{(n)} = \boldsymbol{\tau}^{(n)} + \boldsymbol{\Lambda}^{(n)}\boldsymbol{\mu}_{\mathbf{f}}^{(n)} \qquad (2)$$

If constraints are implied on model 1, we obtain different forms of measurement invariance. The following two sets of equations are important restrictions:

$$\boldsymbol{\Lambda}^{(1)} = \boldsymbol{\Lambda}^{(2)} = \boldsymbol{\Lambda}^{(3)} = \ldots = \boldsymbol{\Lambda}^{(n)} \qquad (3)$$

$$\boldsymbol{\tau}^{(1)} = \boldsymbol{\tau}^{(2)} = \boldsymbol{\tau}^{(3)} = \ldots = \boldsymbol{\tau}^{(n)} \qquad (4)$$

Meredith (1993) has pointed out that there are three forms of invariance that are important for cross-national comparative research: (1) *Configural invariance* which implies that the measurement model, equation 1, holds across all countries; (2) *Metric invariance* which implies that configural invariance holds, as well as equation 3; (3) *Scalar invariance* implies that metric invariance holds as well as equation 4. If configural invariance holds it implies that the measurement instrument is the same across countries, however, comparisons of the measures are still not meaningful. If metric invariance holds, comparisons of relationships between unstandardized measures become meaningful, and if scalar invariance holds it also becomes meaningful to compare the means of the measures.

2.2. Correction for measurement error in measurement invariance testing

It is a well-known fact that the disturbance terms ($\boldsymbol{\delta}$) in the common factor model contain both item specific factors as well as random measurement error (Heise & Borhnstedt, 1970, p.107; Van der Veld & Saris, 2004). This fact is commonly ignored in factor analysis and also in invariance testing, except for Saris & Gallhofer (2007) and Millsap & Meredith (2007), as a result the wrong parameters are estimated and tested. In order to solve this issue, we should separate the random error component from the unique component. The model which enables us to do this is explained in Saris & Gallhofer (2007) and Van der Veld (2006). They make a distinction between two response processes as a result of a stimulus, i.e. the survey item. The first process results in an attitude/opinion or a trait/state (e.g. Steyer & Schmitt, 1990), while the second process results in a response. The processes are represented by respectively equation 5a and 5b:

$$\mathbf{s}^{(n)} = \boldsymbol{\tau}_{\mathbf{s}}{}^{(n)} + \mathbf{C}^{(n)}\mathbf{f}^{(n)} + \mathbf{u}^{(n)} \tag{5a}$$

$$\mathbf{y}^{(n)} = \boldsymbol{\tau}_{\mathbf{y}}{}^{(n)} + \mathbf{Q}^{(n)}\mathbf{s}^{(n)} + \mathbf{e}^{(n)} \tag{5b}$$

In equation 5a **f** is a vector of common factors and **s** a vector of item specific vectors, **u** is a vector of unique components, and $\boldsymbol{\tau}_{\mathbf{s}}$ is a vector of intercepts of the item specific factors. The distinction, *common factor* versus *item specific factors*, was also made by Saris & Gallhofer (2007), following the footsteps of Filmer Northrop (1939, p. 82) and Hubert Blalock (1968). They refer to the common factor as a measure of a concept-by-postulation and to the item specific factor as a measure of a concept-by-intuition. Examples of concepts-by-intuition are ‘judgments’, e.g. *do you like the house you live in*, or ‘feelings’, e.g. *taking all things together, how happy would you say you are*. Thus, concepts-by-intuition are measured with single survey items and their meaning is obvious from formulation. Examples of concepts-by-postulation might include ‘generalized social trust’, or ‘perceived control over one’s life’. A single survey item cannot present generalized social trust or perceived control, but several concepts-by-intuition can form a concept-by-postulation. The difference between a concept-by-postulation (f) and a concept-by-intuition (s) is defined by the model (equation 5a) as the unique component (u). The matrix **C** is a matrix with consistency coefficients, representing the agreement between a concept-by-postulation (f) and a concept-by-intuition (s). We have called these parameters consistency coefficients following Saris & Gallhofer (2007), however, Heise & Borhnstedt (1970, p.107) have referred to these parameters as validity coefficients. The reasoning behind the latter definition is that the larger the coefficient, the better that item specific factor (s) represents the concept-by-postulation (f). We, however, prefer the term consistency coefficient in order to make a distinction with the indicator validity coefficients in multitrait multimethod models, which we will refer to later in this chapter.

In equation 5b **y** is a vector of observed variables, **e** is a vector of random measurement error components, and $\boldsymbol{\tau}_\mathbf{y}$ is vector of intercepts of the observed variables. Furthermore, **Q** is a matrix of quality coefficients, indicating the quality of each observed y.

In this model, equation 5a and 5b, it is assumed that the random error components (e) are unrelated among themselves as well as with the unique components (u), the item specific factors (s), and the common factors (f). The unique component (u) are also uncorrelated among themselves, uncorrelated with the item specific factors (s), and uncorrelated with the common factors (f). Furthermore, the unique components (u) and random error components (e) have a mean of zero.

Next to random measurement error (e), there could also be systematic measurement error (m) as a result of using the same measurement procedure for indicators in the model (Andrews, 1984; Saris & Andrews, 1991; Scherpenzeel & Saris, 1997). This will result in common variance between the indicators due to the common measurement procedure. To put it differently, part of the variance of the common factor (F) could actually be the result of the respondents' systematic reactions to a common measurement procedure. In order to correct for this we will introduce a common method factor in equation 5b.

$$\mathbf{s}^{(n)} = \boldsymbol{\tau}_\mathbf{s}^{(n)} + \mathbf{C}^{(n)}\mathbf{f}^{(n)} + \mathbf{u}^{(n)} \qquad (6a)$$

$$\mathbf{y}^{(n)} = \boldsymbol{\tau}_\mathrm{y}^{(n)} + \mathbf{Q}^{(n)}\mathbf{s}^{(n)} + \mathbf{I}^{(n)}\mathbf{m}^{(n)} + \mathbf{e}^{(n)} \qquad (6b)$$

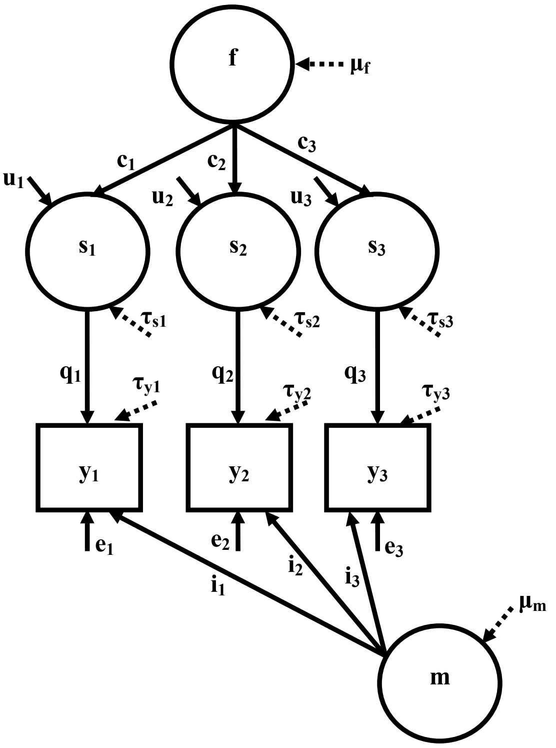
Where **m** is a vector of common method factors that causes the common variance due to the measurement procedure. The matrix **I** contains the invalidity coefficients, which are called this way, because they represent the effect of the common method factor (m) on the indicators. We make the same assumptions as for equations 5a and 5b. In addition, we assume that the common method factors (m) are not correlated with the common factors (f), nor with the item specific factors (s). Furthermore, we assume that the unique component (u) and the random error components (e) are also uncorrelated with the common method factors (m). If these assumptions hold, then the expected value of **y** can be expressed as:

$$\boldsymbol{\mu}_{\mathbf{y}}^{(n)} = \boldsymbol{\tau}_{\mathbf{y}}^{(n)} + \mathbf{Q}^{(n)}[\boldsymbol{\tau}_{\mathbf{s}}^{(n)} + \mathbf{C}\boldsymbol{\mu}_{\mathbf{F}}^{(n)}] + \mathbf{I}\boldsymbol{\mu}_{\mathbf{m}}^{(n)} \qquad (7)$$

Where $\boldsymbol{\mu}_{\mathbf{f}}$ is a vector with the means of the common factors (f) and $\boldsymbol{\mu}_{\mathbf{m}}$ is a vector of means of the method factors (m). All other parameters were introduced and explained previously.

It is not easy to visualize the path model implied by equation 6a, 6b, and 7. In order to clarify the model we have inserted a path model (figure 1) in agreement with these equations for 3 observed variables that measure a common construct. This path model illustrates that the covariance between the observed variables is explained by a common substantive factor f and a method factor m (systematic error). The variances of the item specific factors are explained by the common factor and the unique components. The variances of the observed variables are explained by the item specific factors, the method factor, and a random error component. The mean structure is presented using arrows with the dotted lines. Finally, the model contains several elements with newly developed names, e.g. consistency coefficient, quality coefficient, item specific factor, and also names that are sometimes used in a different context, e.g. unique component. Unfortunately there is no common agreed name for most of the elements in this model, except that f is clearly a common factor, and the y's are clearly indicators. Most problematic in this respect is probably the item specific factor (s). Its name is derived from the fact that those factors only load on a single indicator and are therefore specific for the item that measures this indicator.

Figure 1: Path model of the model represented by equations 6a and 6b.



2.2.1. The 'new' metric invariance test

The constraints to test for metric and scalar measurement invariance are on different parameters compared to the standard procedure. Metric invariance is assessed by testing:

$$\mathbf{C}^{(1)} = \mathbf{C}^{(2)} = \mathbf{C}^{(3)} = \ldots = \mathbf{C}^{(n)} \tag{8}$$

For the reader it will not be immediately clear that this is a test of metric invariance. The essence of a test for metric invariance is that the common factor is expressed on the same scale, i.e. in the same metric, cross-nationally. In any factor model the scales of the latent variables are undefined. The common way to provide a scale for the latent variables is by fixing the loading of one of the indicators to 1 (unit loading identification). The result is that the latent variable is expressed on the same scale as the response scale that was used to measure that indicator. This principle ensures that our metric invariance test, equation 8, is truly a test of metric invariance. We will deal with identification later, but we have to introduce this topic here already, less detailed though, to illustrate that this is true. The scales for the item specific factors (s) are defined by fixing the quality coefficients to 1. By doing that, the item specific factors are expressed in the same metric as the observed variables. Furthermore, the scale of the common factor is defined by fixing one of the consistency coefficients to 1. So, now the metric of the common factor is the same as the metric of the observed variable for which both the quality and consistency coefficient are 1. If then equation 8 holds, then the metric of the common factor is the same across all countries. Under the condition of course, that the indicators are observed with the same response scale in all countries.

The meaning of this constraint follows from the interpretation of the consistency coefficient, which is the agreement between what we intended to measure (f) and what we measured (s) after correction for measurement error. To put it differently, the consistency coefficient indicates how well the item is *understood*, in the light of what we intended to measure with the item. If the item is *understood* in the same way as the intended measure (f), then the consistency coefficient is perfect, i.e. 1. Hence, cross-national equality constraints on the consistency coefficients imply that the item is *understood* in the same way, i.e. *conveys the same meaning (*Kumata & Schramm, 1956), across countries.

If this constraint, equation 8, also results in metric invariance, what then is the difference with the standard test? If we express the standard test for metric invariance (equation 3) in terms of the parameters from equation 6a and 6b, the test would look as follows: $(\mathbf{CQ})^{(1)}=(\mathbf{CQ})^{(2)}=\ldots=(\mathbf{CQ})^{(n)}$. Thus in the commonly used procedure it is assumed that the product of the measurement quality (q) and the consistency (c) is the same across countries. That is an assumption which is not warranted. Several studies have shown that the quality of measures varies across Europe, e.g. Scherpenzeel (1995b) for life satisfaction, Saris & Gallhofer (2003) for a variety of measures in the ESS. Given the evidence that the quality is in general not equal across countries, we should impose equality only on the consistency coefficients. This is exactly what is done in the test (8) we propose.

2.2.2. The 'new' scalar invariance test

For the test of scalar invariance there are similar issues, which leads to the following model restriction for the scalar invariance test (in addition to the metric invariance constraint):

$$\boldsymbol{\tau}_{\mathbf{s}}^{(1)} = \boldsymbol{\tau}_{\mathbf{s}}^{(2)} = \boldsymbol{\tau}_{\mathbf{s}}^{(3)} = \ldots = \boldsymbol{\tau}_{\mathbf{s}}^{(n)} \qquad (9)$$

If this equality holds then the common factor has the same zero point across all countries, and thus it becomes meaningful to compare latent scores across countries. If we express the standard test for scalar invariance (equation 4) in terms of the parameters from equation 6a and 6b, the test would look like: $(\boldsymbol{\tau}_y+\mathbf{Q}\boldsymbol{\tau}_s)^{(1)}=(\boldsymbol{\tau}_y+\mathbf{Q}\boldsymbol{\tau}_s)^{(2)}=\ldots\ldots=(\boldsymbol{\tau}_y+\mathbf{Q}\boldsymbol{\tau}_s)^{(n)}$. In this equation the intercepts of the indicators (τ_y) and the measurement quality (q) can vary across countries due to the measurement procedure. Cross-national variation in the quality was already discussed in the metric invariance test. Cross-national variation of the intercept of the indicators (τ_y) has the same roots. An intercept commonly changes with the addition of extra predictors. In equation 6b there is an extra predictor (compared to equation 1) for the indicator, namely the method factor (m). If the method factor has a mean different from zero and if this mean varies cross-nationally, than the intercept of the indicator (τ_y) will also vary. For this reason we suggest to test the restriction of equation (9) and not the restriction specified in equation (4)

2.3. Strategy for measurement invariance testing

There are two general approaches to testing for invariance, the top-down approach and the bottom-up approach. The top-down approach starts with the most constrained model, in our case this is the model with equal loadings and equal intercepts. In addition to the measurement invariance constraints, the factor model itself also constraints cross-loadings and correlated error terms at zero. All the constraints are tested. If the model fits, according to some criterion, there is no problem. However, if the model is rejected *according to some criterion*, improvements can be made to the model by releasing constraints. The big question is: where to start? The number of constraints is very large. For example in a simple single factor model with 4 indicators scalar invariance across 20 countries results in 280 constrained parameters. Which ones should we release? Do we first introduce correlated errors and cross-loadings, or do we first release the measurement invariance constraints. Therefore, a good reason not to start with the most constrained model is that one immediately starts with a huge number of constrained parameters that can all potentially be incorrect. Another reason not to start with the most constrained model is that measurement invariance constraints can cause residual covariances between the items in some countries. These residuals might be significant and one therefore might want to introduce (estimate) those correlated errors, but that would be a mistake because they are artifact of the measurement invariance constraints.

A better approach is the bottom-up approach, where one starts with the least constrained model, i.e. configural invariance, and then proceeds by introducing more constraints to the model. The advantage of the bottom-up approach is that the problems one faces in each step are more manageable compared to the top-down approach. The bottom-up procedure will be discussed in the following 3 paragraphs dealing with the different forms of invariance testing. Do note, however, that nothing is mentioned about which goodness-of-fit measures are used. Instead we will use the phrase '*according to some criterion*'. In the section 'Introducing an alternative model test procedure' we will come back at this issue.

2.3.1. The configural invariance test

The test for configural invariance is in essence a test to check whether the indicators measure the latent variable(s) they are intended to measure. It is imperative that there is a test for configural invariance. The reason is that we will introduce constraints in the model during the phase of testing for metric and scalar invariance. When those models show a lack of fit, we want to be able to uniquely attribute this to the extra constraints imposed on the model by metric and scalar invariance. That is not possible if the less constrained model is not tested; hence a test for configural invariance is imperative.

In the social sciences, measurement instruments often only have 1, 2, or 3 indicators. Such measurement instruments cannot be tested; the 1 and the 2 indicator model are not identified (without restrictions) and the 3 indicator model is just identified. So, only in case an instrument has 4 indicators, or more, it is possible to test whether the instrument exhibits configural invariance. When a measurement instrument only has 2 or 3 indicators it is possible to do a test, but that requires that the measurement model is extended with: (1) one or more other measurement instruments, or (2) causes (predictors) and/or consequences of the construct, or (3) extra within country restrictions, or (4) any combination of these possibilities.

What constraints are tested in the test for configural invariance? By definition we test whether the indicators measure the latent variable(s) they should measure. For a model with several latent variables and a set of predictors this implies that we test whether the following constraints hold: (1) the correlations between the unique factors are zero, (2) all cross-loadings are zero unless the theory dictates otherwise, and (3) the predictors have no direct (thus zero) effect on the indicators. If the test indicates that the model is misspecified *according to some criterion*, the misspecified parameter(s) should be estimated, because ignoring these misspecifications could lead to biased parameter estimates. Obviously we would like to understand precisely why this misspecification occurs, but such post-hoc reasoning will only be helpful for future research. For the study at hand one should estimate the misspecified parameter, or any equivalent solution to the misspecification, anyway.

After the model is judged acceptable *according to some criterion*, it is time to select the reference country for the metric (and scalar) invariance test. The test for metric invariance assesses whether the consistency coefficients are equal to each other (9). It is therefore necessary to have a reference country for which the consistency coefficient of each indicator is not too extreme compared to the other countries. In order to find this reference country, one should make a table with a country on each row and the consistency estimates of the indicators in the columns. By sorting the table one can easily find a country which is somewhere in the middle and which has no extreme estimates. This should be the reference country for the metric invariance test. The reason for following this procedure is one wishes to compare as many countries as are available. If a country, with high or low factor loadings compared to the average, is selected as the reference country it is *more likely* that that country is not invariant. If an non-invariant country is selected, it will result in a smaller set of comparable countries, compared to procedure we suggested. The reason for this is that one cannot free parameters of the reference country to become non-invariant, because those parameters are already free. Obviously, this procedure is not flawless, there are specific configurations of countries that would result in the opposite that this procedure tries to accomplish. For example, a configuration with one country at the average and two large groups of countries at the extremes, could lead to a smaller set of comparable countries if the *average* country is selected.

Apart from choosing a reference country, one should also decide on a referent indicator, i.e. an indicator that is used to define the scale for the latent variable and therefore drops out of the test for metric and scalar invariance. In principle the choice of a reference indicator should be made for an indicator which is known to be invariant; but this is something we cannot know. Another strategy would be to use an indicator that has the highest face validity for the concept that we wish to measure. This is our preferred choice, but it should be supported to by an analysis to see whether that loading of that indicator indeed shows little variation across countries. The choice of a non invariant reference indicator can be quite problematic as Johnson, Meade, and DuVernet (2009), Yoon & Millsap (2007), and Rensvold & Cheung (2001) have pointed out.

2.3.2. The metric invariance test

This test will reveal which loadings are non-invariant across countries. If non-invariant loadings are present, one faces the problem where to start releasing the constraints. In principle one should start with the indicators that are most deviant. These indicators are easily found in the table that was created to select the reference country. Model adjustments should be made one-at-a-time until an acceptable model is obtained *according to some criterion*. If model adjustments are necessary, it *could* result in partial metric invariance as described by Byrne, Shavelson, and Muthén (1989). Partial metric invariance means that at least one metric invariant indicators per factor remain, plus the referent indicator which is also assumed invariant. If there is partial invariance then composite scores or sum scores, which are often used in research, should not be used since they will bias substantive conclusions (Saris & Gallhofer, 2007, ch. 16). On the other hand, Byrne, Shavelson, and Muthén (1989) have pointed that when the sources of non-invariance are explicitly modeled, then only one invariant indicator is enough for meaningful cross-national comparisons within the context of SEmodeling. A final strategic note on metric invariance testing is that one should not introduce correlated unique components or correlated random error components, because they should have been detected during the configural invariance testing. If they are found during the metric invariance testing, they should be the result of the restrictions on the parameters implied by metric invariance.

2.3.3. The scalar invariance test

If the test for metric invariance resulted in indicators that lack metric invariance it will make no sense to include those non-invariant indicators in the test for scalar invariance. The reason is that metric invariance is a requirement for scalar invariance. As a consequence, indicators that were found to lack metric invariance should not be included in the constraints for the scalar invariance test. Therefore, for the non-metric-invariant indicators the τ_s should be estimated without constraints. The test of the scalar invariance will indicate which indicators for which countries are not scalar invariant *according to some criterion*. If problematic indicators are present, one faces the problem where to start releasing the scalar invariance constraints. In this case we do not have a list of unconstrained estimates of the intercepts (τ_s) of the true scores, as we did have for the consistency coefficients in the metric invariance test.

The reason is that one cannot estimate all these intercepts without restrictions[1]. Therefore we suggest a different approach, which is to look at the residuals of the means of the observed variables. In principle one should start to free that intercept (τ_s) which has the largest difference from the reference value. Model adjustments should be made one-at-a-time until an acceptable model is obtained *according to some criterion*. If such model adjustments are necessary, it could result in partial scalar invariance as described by Byrne, Shavelson, and Muthén (1989). Partial scalar invariance means that at least two scalar invariant indicators per factor remain. Again, one should be careful with the construction of composite scores if there is partial invariance (Saris & Gallhofer, 2007, ch. 16).

2.4. An alternative model evaluation procedure

2.4.1. The detection of misspecifications

So far we have used the phrase *accep*table *to some criterion* several times without specifying what that criterion is. For these generally complex models it is not clear which criterion to use, commonly a mixture of goodness-of-fit indices are used with the cut-off criteria suggested by Hu and Bentler (1999). Recent studies have, however, shown that fit indices with fixed critical values (e.g., the RMSEA, GFI) don't work as they should, because it is not possible to control for type I and type II errors (Barret, 2006; Marsh et al., 2004; Saris et al., 2009). This means that correct theories are rejected and incorrect theories are accepted in unknown rates. An alternative procedure, *the detection of misspecifications*, has recently been suggested by Saris et al. (2009). The procedure is build upon the idea that models are simplifications of reality and are therefore always misspecified to some extent (Browne & Cudeck, 1993; MacCallum, Browne, and Sugawara 1996). This is normally problematic because when the power of the test is large the model will be rejected even because of irrelevant misspecification. This is normally problematic because when the power of the test is large one can detect even the smallest misspecification. However, in our procedure we can control which magnitude of misspecification should be detected with high power.

The traditional procedure to detect misspecifications is by use of the modification index together with the expected parameter change to judge whether the constrained parameter is a misspecification (Kaplan, 1989; Saris, Satorra, and Sörbom, 1987). However, the modification index (MI) is sensitive to the power of the test (Saris et al., 2009), therefore one should take the power of the MI-test into account. The power of the MI-test to detect a misspecification of size delta or larger for any constrained parameter can be derived if the non-centrality parameter for the MI-test is known. The non-centrality parameter (ncp) can be computed with the following formula:

$$\text{ncp} = (\text{MI}/\text{EPC}^2)\ \delta^2 \qquad (10)$$

In this equation the MI is the modification index and the EPC is the expected parameter change which both can be found in the output that SEM software produce (for more details see Saris et al., 2009). Furthermore, δ (delta) is the size (magnitude) of the misspecification we would like to detect with high power. Its value can be chosen by the researcher and may vary across disciplines and the state of the theory under investigation. However, guidelines exist as to which magnitude of misspecifications are important to be detected under general conditions. These suggestions will be discussed later. The power of the test, for which we want to control in the end, can be obtained from the tables of the non-central χ^2-distribution.

Whether or not a constrained parameter is a misspecification is judged from the combination of the power, which can be high or low, and the modification index, which can be significant or not. The decision rules are presented in table 1. There is a misspecification when the power to detect a misspecification of delta is low and the modification index is significant. There could also be a misspecification when the power is high and the modification index is significant. In that case, the MI could be significant due to the high power or because there is a large misspecification. Thus, when in this instance the expected parameter change is larger than delta, we decide that there is a misspecification. Other combinations are also possible and indicate either no misspecification or a lack of power to detect a misspecification. One can imagine that this procedure is quite laborious because for each constrained parameter we have to compute the power of the test and then decide on a misspecification using the judgment rules in table 1. For a simple multi-group factor model with 4 indicators, 20 countries and the scalar invariance constraints, there are already 280 constraints. Therefore a software program called JRule[2] has been developed by Van der Veld, et al. (2008) which automates the whole procedure.

Table 1: The judgment rules.

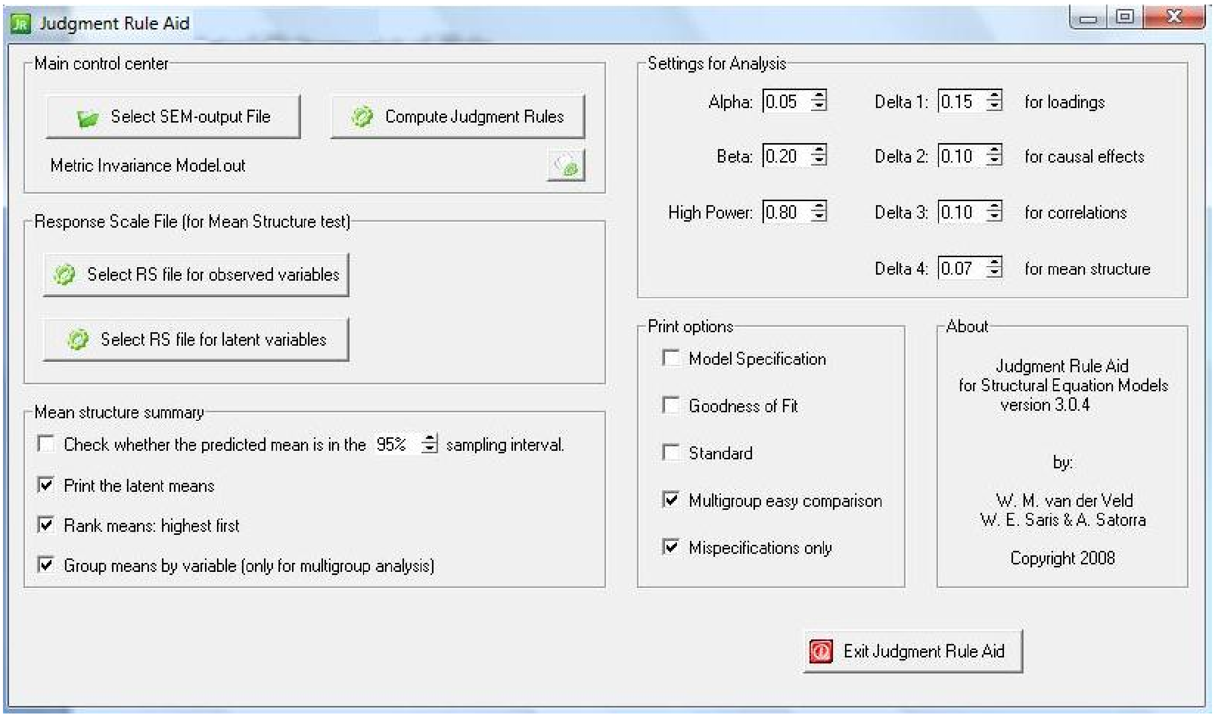
Judgment Rule			
1	*MI = not significant*	*power = high*	**No misspecification**
2	*MI = significant*	*power = low*	**Misspecification present**
3	*MI = significant*	*power = high*	**Use EPC**
4	*MI = not significant*	*power = low*	**No decision**

2.4.2. Using JRule in cross-national analysis

JRule reads the output of either LISREL (Jöreskog & Sörbom, 1996) or Mplus (Muthén & Muthén, 2007) and collects model information (MI & EPC) necessary to make a judgment about whether the constrained parameters in the model are misspecified or not, or whether there is no statistical basis to make a sound judgment. In figure 2 a screenshot of JRule is presented. The program is fairly simple to use. The user only has to select a SEM-output file and then press the button 'Compute Judgment Rules'. JRule will then compute all judgment rules for the constrained parameters and present them in a text file, readable with any text editor. One can furthermore see that the user can specify the levels of α and β for the test. In addition, the user can specify the magnitude of the misspecification that he likes to detect, or better, that he does not want to ignore.

A special feature – Multigroup easy comparison – makes the evaluation of multi-group models a lot easier than it would be when the standard SEM output is used to make an evaluation. The reason for this is that SEM programs commonly produce output group by group, instead of parameter by parameter. In case of measurement invariance testing, the interest is especially in detecting whether there are countries that are deviant. Already in simple models one gets lost in the amount of output, because it is not structured efficiently for this purpose.

Figure 2: Screenshot of JRule.



Main control center

Metric Invariance Model.out

Response Scale File (for Mean Structure test)

Mean structure summary

☐ Check whether the predicted mean is in the 95% sampling interval.

☑ Print the latent means

☑ Rank means: highest first

☑ Group means by variable (only for multigroup analysis)

Settings for Analysis

Alpha: 0.05 Delta 1: 0.15 for loadings

Beta: 0.20 Delta 2: 0.10 for causal effects

High Power: 0.80 Delta 3: 0.10 for correlations

Delta 4: 0.07 for mean structure

Print options

☐ Model Specification

☐ Goodness of Fit

☐ Standard

☑ Multigroup easy comparison

☑ Mispecifications only

About

Judgment Rule Aid
for Structural Equation Models
version 3.0.4

by:

W. M. van der Veld
W. E. Saris & A. Satorra

Copyright 2008

The output JRule produces is organized in such a way that it becomes very easy to compare the same constrained parameter across all countries. Figure 3, shows a fragment of such an output. In that figure an overview is presented for the misspecifications in a part of the variance-covariance matrix of random error components. So, this overview indicates whether and where there are misspecified correlated errors. On each row, one can find the judgment rules for a single parameter for all countries in the analysis. Here the countries are indicated with G1 to G10. So, one can see that the parameter ‘X2 with X1’ is misspecified (JR=2) in G2, G3, G7, in addition for 6 countries one lacks statistical information (JR=4) to make a sound judgment of whether that parameter is misspecified. It is clear that organizing the output in this way makes it a lot easier to see what parameters are misspecified in which countries, and also whether a certain parameter is misspecified in many countries, or only incidentally. In addition, JRule also provides the opportunity to print only the misspecified parameters, which makes it even easier to evaluate whether and where something is wrong.

Figure 3: Fragment from a JRule output with Multi-group easy comparison, indicating whether and where there are misspecified correlated errors[a].

RELATION	BETWEEN	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
X1	X1	-	-	-	-	-	-	-	-	-	-
X2	X1	4	2	2	4	1	4	2	1	4	4
X3	X1	2	3	3	1	3	4	4	3	3	2
X4	X1	1	1	3	1	3	1	1	1	3	1
X5	X1	3	1	3	1	3	1	1	1	1	1
X2	X2	-	-	-	-	-	-	-	-	-	-
X3	X2	2	1	1	1	3	4	4	3	3	4
X4	X2	1	1	1	1	1	1	1	1	3	1
X5	X2	3	1	1	1	3	1	1	3	1	1

[a] *G1, G2, G3, refer to the groups or countries in the analysis. JRule presents a table in the output which makes the link between the names the user has given to the groups and G1, G2, et cetera, explicit. These short names, G1, G2, et cetera, are used to enable a more comprehendible lay-out of the results. The numbers in the table are judgment rules, which must be interpreted using table 1.*

3. The cross-national comparability of GST

3.1. Theoretical context

GST is a central variable in the social sciences because it is assumed to be fundamental for a healthy society (Putnam, 1993; Fukuyama, 1995). Nevertheless it is not clear how, because *trust appears to work somewhat mysteriously* (Uslaner, 2000, p. 569). That is, it is not clear why some people are more trusting than others, or why citizens in some countries are more trusting than in other countries. We are interested in explaining GST. There are two categories of explanations, country level explanations and individual level explanations. We focus on the individual level explanations. Delhey & Newton (2003) mention five individual level theories which they have named: Personality theory, Success and well-being theory, Voluntary organization theory, Social network theory, and Community theory. In the personality theory it is assumed that it is attitudes formed in early-childhood and personality features that determines trust in others (Erikson, 1950). In the success and well-being theory it is assumed that those who have more success in life, i.e. those who are better off, will be more trusting of others (Putnam, 2000). The Social network theory suggests that direct participation in the social networks of everyday life, e.g. friends, family, and colleagues, will promote trust in others. The Community theory suggests that trust in others is fostered by features of the local context in which people live. This is in contrast to what Delhey & Newton (2003) named Societal theory which is concerned with country level variables such as GDP or democratic level, which we ignore in our study. Finally, the Voluntary organizations theory suggests that it is participation in formal organizations which promote trust in others. This theory is one of the more popular explanations. It has been put forward by Putnam (1993), who, following Alexis de Tocqueville and John Stuart Mill, argues that participation in civic organizations leads people to trust each other. There is, however, no consistent evidence for this idea (Torcal & Montero, 1999, Delhey & Newton, 2005; Gesthuizen, Scheepers, Van der Veld, Völker, in press; Saris & Gallhofer, 2007, ch. 15) and it has also been criticized on theoretical grounds (Levi, 1996; Newton, 1997). We agree with these critics and therefore will not further develop this idea here. Finally, there are three often used demographic variables (gender, age, and education) which also have an effect on GST, but the theoretical justifications are not well-founded.

In their study, which tests the 5 theories (Personality, Success and well-being, Voluntary organization, Social network, and Community), Delhey & Newton (2003) find some hard to explain results. The most eye-catching result is that the relative importance of these theories differs across countries. This implies that people living in these countries trust for different reasons. This is hard to believe. People are people so why should it be that in country A happy people trust others more, while in country B more social active people trust others more. One would expect that the same causes, with more or less the same effect, are at work in country A as well as in country B. After all we are dealing with the same causal mechanisms. The fact that Delhey and Newton (2003), as well as other scholars (e.g. Kaasa & Parts, 2008; Zmerli & Newton, 2008) find these inconsistent results might be that they do not correct for measurement error. For example, it is more than likely that instruments used to measure the causes (and consequence, i.e. GST) differ in quality across countries. If the ‘true’ effects of the causes are the same across countries, then we will find that the estimated effects will be different, due to attenuation for measurement error.

Uslaner (2000) provides another explanation for GST which does not fit into any of the previously mentioned theories. He argues that the trust in others is explained by how religious one is. Uslaner (2000) finds that Christian fundamentalists are substantially less likely than other believers to say that they trust other people. The rationale behind this is based upon the perception that Christian fundamentalists do take part in civic life, but only with their own kind. As a result they trust those who are similar, but not the general others. We will refer to this theory as the Orthodoxy theory.

The aim of this last part of the chapter is to test these theories of the causes of GST. We will do this using the procedures we suggested in this chapter, i.e. correcting our measures for measurement error and using a different model evaluation procedure.

3.2. Data

The data for the analysis are collected by the ESS, and we use round 1 data. The ESS is built upon the belief that cross-national comparative research requires more than just having respondents completing the same questionnaire in different countries. The procedures, used in the ESS, to ensure cross national equivalence are pretty elaborate and involve among others the control of sampling designs, questionnaire design, translation procedures, data entry, and multitrait multimethod (MTMM) experiments for quality control. The whole survey process is being controlled as much as possible in every participating country. GST is measured in the ESS with three survey questions. The formulations are presented in table 2.

Table 2: The formulation of the indicators of GST[a].

Names	**Formulation of survey items**
Trust	Using this card, generally speaking, would you say that most people can be trusted, or that you can't be too careful in dealing with people? Please tell me on a score of 0 to 10, where 0 means you can't be too careful and 10 means that most people can be trusted.
Fair	Do you think that most people would try to take advantage of you if they got a chance or would they try to be fair?
Help	Would you say that most of the time people try to be helpful or are they mostly looking out for themselves?

[a] *Measured on an 11 point scale running from 0 to 10 with item specific endpoint labels.*

In order to test the theories (Personality theory, Success and well-being theory, Social network theory, Community theory, and Orthodoxy theory) we need indicators for these theories. There is however little agreement about which variables are important and even less how they should be measured. In addition we are limited in our possibilities to what is available in the ESS. We have used one measure for most theories, and selected those indicators that closely resembled the indicators used in the study by Delhey & Newton (2003).

The ESS, round 1, does not provide any indicators for the Personality theory, so we ignore this theory in our analysis. The success and well-being theory was tested with two measures, an objective and a subjective one. The subjective indicator is an item about how happy one is (Happy) and the objective indicator is an item about whether people find it difficult to make ends meet (MeetEnd). The social network theory is tested using a measure that asks for frequency with which ones meets with friends, relatives or work colleagues (Social). The community theory was tested with two measures, an objective and a subjective one. The subjective measure is about whether one feels safe at night in the neighborhood (FlSafe) and the objective measure is about the size of the city where one lives (Urban). The Orthodoxy theory was tested with a measure about how important religion is in ones life (Religs). The formulations of these measures are presented in table 3. We have skipped the formulation of the measures for gender and age, since they are standard measures.

Table 3: The formulation of the predictors of GST.

Names	**Formulation of survey items**
Happy[a]	Taking all things together, how happy would you say you are?
MeetEnd	How do you feel about your household's income nowadays? (1) Living comfortably on present income, (2) Coping on present income, (3) Finding it difficult on present income, (4) Finding it very difficult on present income.
Social[b]	How often do you meet socially with friends, relatives or work colleagues?
FlSafe	How safe do you – or would you – feel walking alone in this area after dark? (1) Very unsafe, (2) Unsafe, (3) Safe, (4) Very safe.
Religs[c]	How important religion in your life?
Urban	Which phrase on this card best describes the area where you live? (1) A big city, (2) The suburbs or outskirts of a big city, (3) A town or a small city, (4) A country village, (5) A farm or home in the countryside.
Educ[d]	What is the highest level of education you have achieved?

[a] *Measured on an 11 point scale with endpoint labels (0=Extremely unhappy, 10=Extremely happy).*

[b] *Measured on a 7 point fully labeled scale with subjective frequencies (never……everyday).*

[c] *Measured on an 11 point scale with endpoint labels (0=Extremely unimportant,10= Extremely important).*

[d] *Measured with country specific scales, lower values mean lower education. Basically incomparable values due to differences in educational systems.*

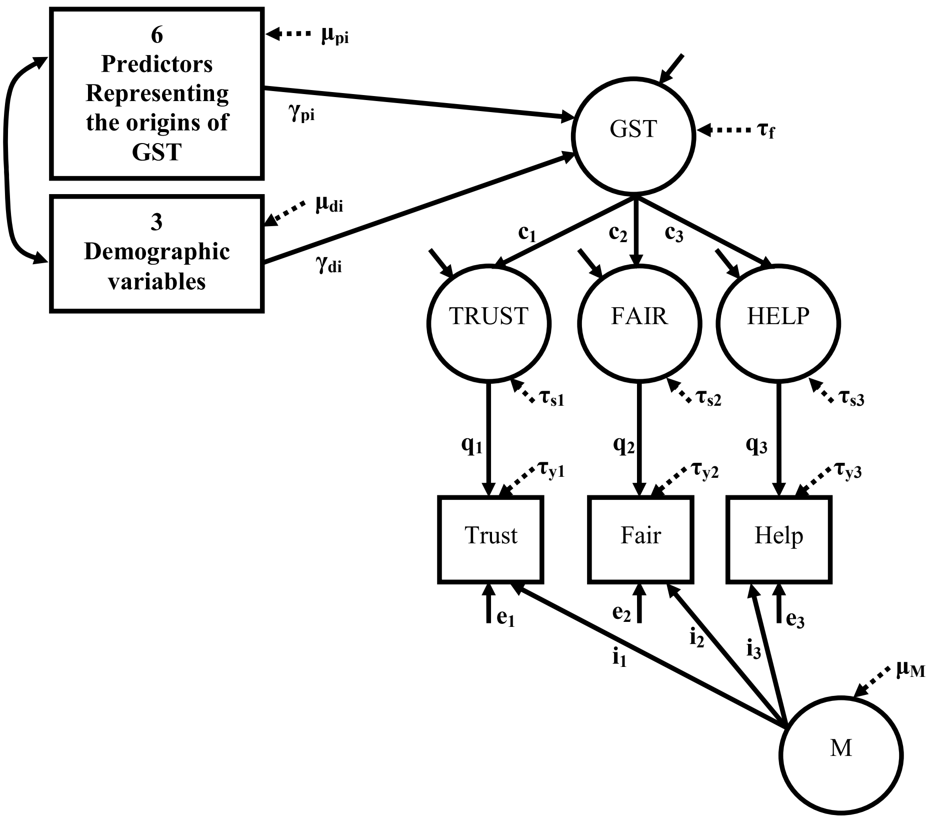
3.3. Model identification

The model we test is presented in figure 4. This model is not identified without restrictions. In order to define the scales for the item specific factors (s), all quality coefficients were fixed to 1. Differences in the quality are still possible because the random error components (e) are not constrained. The scale for GST is defined by fixing the consistency coefficient of TRUST to 1. Furthermore, the scale for the systematic measurement error factor (m) is defined by fixing all invalidity coefficients to 1. The result is that the method factor (m) has the same effect within a country on the different items, but can have a different effect in each country because the variance of the method factor (m) can differ across countries. Corten, Saris, Coenders, Van der Veld, Aalberts, and Kornelis (2002) have studied which specification of the effect of the method factor works best. They found that a specification where the method factor had an additive scale dependent effect fitted the data best. We have used that specific specification.

Defining the scales for the latent variables, however, does not make the model identified. There is still an identification problem in the measurement part of the model. This problem is found in the random error component (e) and method factor (m) variance. There are no equality restrictions possible to make this model identified. There are nevertheless two solutions to this problem. A first possibility is to extend the model and data collection design following the multitrait multimethod (MTMM) approach (Scherpenzeel, 1995a; Saris & Andrews, 1991). That would, however, make the specification, testing and estimation of the model very complex. A second, more simple, solution is to fix the values of the random error components and method factor variances to a reasonable value. In that case we do not have to estimate those coefficients, which makes the model identified. But what are reasonable values for the random and systematic error components? We can obtain reasonable values in two ways. A first possibility is to predict the random and systematic error components using the Survey Quality Predictor (Saris & Gallhofer, 2007; Saris, Van der Veld, and Gallhofer, 2004). Unfortunately, currently that procedure only works for Dutch, English, and German questionnaires. The second possibility, and our choice, would be to estimate the random (e) and systematic error (m) components in a different context and then introduce these estimates as fixed values in the model (figure 4). This is possible because the items Trust, Fair, and Help have been collected as part of a multitrait multimethod (MTMM) experiment in the

ESS. This design enables us to estimate the random and systematic error components using the MTMM approach.

Figure 4: Path model[a] to test theory of the causes of GST.



[a] *The means of the predictors is indicated by μ_{pi}, where the subscript pi refers to different predictors in the model. The regression effects of the predictors are indicated with γ_{pi}. In the model it looks as if there is only one regression effect, but in fact there is one for each predictor. The means of the demographic variables is indicated by μ_{di}, where the subscript di refers to the different demographical variables in the model. For the regression effects of the demographic variables (γ_{di}) it's the same story as for the predictors. In addition, demographic variables are in a sense also predictors, but because there is no theory we have made this distinction.*

After these restrictions, the model is identified and a test can be performed for configural and metric invariance. A test for scalar invariance is however not possible, because the means of the method factors are not identified. We need extra restrictions to identify the model. Our solution is to assume that the method factor means have the same value in all countries.

Because the actual value is irrelevant, we have fixed the method factor means to zero. This assumption might not be warranted, however, there is no easy alternative[3].

Even after all these restrictions, a test for scalar invariance cannot be performed because the intercepts of the item specific factors (τ_s) and the intercepts of the observed variables (τ_y) cannot be simultaneously estimated. We need another restriction to identify the model. Byrne and Stewart (2006) have suggested to fix the intercepts of the first order factors, in our case the item specific factors (s), to 0. But their model is slightly different from our model, in that their model has multiple indicators for each first order factor. In our case there is only one indicator for each first order factor. It is therefore arbitrary whether we would fix the intercepts of the item specific factors (s) or of the observed variables (indictors) to zero. We have chosen for the latter option.

3.3.1. The estimation of random and systematic error variance (quality and invalidity)

In order to obtain the estimates for the random error and method variances for all countries we first estimated an MTMM model for each country. We will not discuss this model in any detail, except that the model that we estimated was the classic MTMM model as used by Andrews (1984) and described by Batista-Foguet & Coenders (2000), but using a multi-group design called SB-MTMM (Saris, Satorra, and Coenders, 2004) to minimize response burden. The first round of the ESS contains SB-MTMM experiments for Trust, Fair, and Help, that we have used to estimate the parameters of interest. We used the full information maximum likelihood procedure available in LISREL 8.8 (Jöreskog and Sörbom, 1996) to estimate our model and account for missing data[4]. The results of the analysis are presented in table 4. The 2nd to the 4th column contain the estimates of the random error components. One can see that the estimates are significantly different from zero, while zero would indicate a measure without random measurement error. In addition, the random error components vary across different items as well as across countries. The latter result justifies that we make a distinction between the unique components and random error components as we did in equation 6a and 6b. Column 5 presents the systematic error components, the method factor variances. In this specific case the systematic error components are not significant in most countries, this is indicated by either *ns* or FI in table 4. There is a good reason why the systematic error components are not significant in most countries. The main reason is that this scale was tested in the ESS pilot study for round 1 and the results showed that the method effects, i.e. systematic measurement error, were not significant for this scale. A more theoretical reason is that the response scales are item specific and thus reduce the common variance due to a systematic reaction within respondents on the response format. This format is in sharp contrast to response scale formats such as agree-disagree, or never-often. Saris, Revilla, Krosnick, and Schaeffer (in press) have studied these differences and found that item specific response scales perform much better, i.e. little systematic measurement error, compared to agree-disagree scales.

The values of the random error components in table 4 are the ones that we have introduced as fixed values in the complete model as depicted in figure 4. For the systematic error components we have chosen for a much simpler solution. For most countries we do not find significant systematic measurement errors, therefore we will ignore the method factor in the tests for measurement invariance.

3.4. Two unfortunate facts

Our aim is to make a cross-national comparison of several theories that explain GST. This would require that, (1) the measures of all variables in our model show scalar measurement invariance and (2) that we can correct all our measures for measurement error. This is possible for the endogenous variable GST, but not for the exogenous variables. It is not absolutely necessary that the exogenous variables are scalar invariant. If they are not, we could still determine which theory is the most important theory within each country, and those results can be compared cross-nationally. In contrast we cannot say, in case of the absence of scalar invariance, that ‘success-and-well-being’ has twice as much explanatory power in the UK compared to Italy. However, we would already be happy when we can compare the importance of each theory within countries. That, however, is not possible either. The reason is that in path (and regression) analysis it is assumed that the exogenous variables are observed without measurement error. If this is not the case, and that’s very likely, then the estimates of the path-coefficients are biased. An alternative would be that we correct for measurement error in a similar way as we did for GST. Unfortunately, this is not possible with the data that we have. We do not have MTMM data available in the ESS for the exogenous variables. Another alternative would be to use the multiple indicators approach (Ganzeboom, 2009) for each exogenous variable. That is also no option, because we lack multiple indicators. The bitter conclusion is then the we can estimate all the paths, but we cannot – with any confidence – interpret the results due to the distorting effect of measurement error.

Table 4: Estimates of the variance of random and systematic error components per country[a].

	Random error variance			Systematic error variance
Country[b]	***Trust***	***Fair***	***Help***	**Method**
Sweden	1.44	1.53	1.83	0.01[FI]
Austria	1.32	1.24	1.17	0.01[FI]
Belgium	1.29	1.75	1.24	0.24[ns]
Switzerland	1.69	1.21	1.71	0.01[FI]
Czech Republic	1.48	1.37	1.40	0.61
Germany	1.99	1.84	1.91	0.01[FI]
Denmark	0.86	0.90	1.62	0.01[FI]
Spain	0.93	1.31	2.03	0.24[ns]
Finland	1.30	1.27	1.73	0.41[ns]
Great Britain	1.46	1.28	1.70	0.01[FI]
Greece	0.89	0.95	1.28	0.01[FI]
Ireland	1.61	1.28	1.64	0.01[FI]
Israel	2.27	1.69	2.62	0.44[ns]
Italy	1.66	2.14	2.28	0.50
Netherlands	1.10	1.13	1.20	0.01[FI]
Norway	0.86	0.67	2.26	0.01[FI]
Poland	1.76	2.15	2.72	0.63
Portugal	0.78	0.92	0.96	0.01[FI]
Slovenia	2.65	2.32	2.09	0.55[ns]

[a] *All estimates are significant at $\alpha=0.05$, unless stated otherwise. ns denotes that the estimate is not significant. FI denotes that the parameter is fixed to 0.01. This was necessary if the estimated variance was negative, in all instances where this happened the estimate was not significant.*

[b] *Nineteen countries are presented, while there are 22 countries that participated in the first round of the ESS. The difference is due to the fact that in France, Hungary, and Luxembourg no MTMM experiments were conducted or ill-conducted.*

The question that pops up is then whether we should have introduced the theory about the causes of GST. We do have two reasons to introduce such a theory anyway. Firstly, in the common approach (see references in the first sentence of this chapter) the consequences of the presence of measurement error as well as the comparability of the measures are ignored. We explicitly want to illustrate that one should not ignore these issues by confronting the reader with a theoretical context first and then stress the requirements necessary to perform a test on the theories. Secondly, there is another use for the exogenous variables too. We can use them to over-identify the model so that we have a test of configural invariance. This is an important test, as discussed in the section *strategy for configural invariance testing*, because the measurement model for GST is only exactly identified after introduction of the quality and invalidity coefficients in the model as fixed parameters (see the section on *model identification*). Thus, by including exogenous variables in the model we obtain a test for configural invariance. It is for these reasons that we decided to include a section on the theoretical context, i.e. the causes of GST. We will provide the results, i.e. the effects from the exogenous variables on GST and also interpret them. However, because we cannot deny the possibility that there is measurement error in the predictors leading to biased substantive conclusion, the conclusions should be taken with a grain of salt.

3.5. Model estimation

The model parameters are estimated with the robust maximum likelihood estimation procedure (Satorra & Bentler, 1994) in LISREL 8.8 (Jöreskog & Sörbom, 1996), using the asymptotic covariance matrix to correct for non-normality in the observed data. Missing values were dealt with using listwise deletion. Full Information Maximum Likelihood (FIML) shows a superior performance (Enders & Bandalos, 2001) compared to listwise deletion, however FIML is incompatible with robust ML. Listwise deletion produces unbiased estimates under MCAR conditions and is not efficient under MCAR/MAR because cases are deleted that do have observed values on some variables. We can live with this loss of efficiency because the number of cases remains sufficient for our analysis. Furthermore, we have used the design weight, present in the ESS data, to correct for cross-national differences in the sampling procedures.

Finally, all models are evaluated using the procedure suggested by Saris et al. (2009) as implemented in JRule (Van der Veld et al., 2008). For the evaluation, i.e. the computation of the judgment rules, we have used the following settings: α=0.05, high power=0.80, δ_1=0.30, δ_2=0.15, δ_3=0.15, and δ_4=0.07.

3.5.1. Results for configural invariance

For the test of configural invariance we can ignore the mean structure, thus all variables are expressed in deviation scores for this analysis. We estimated the model for all 19 countries and then analyzed the output with JRule. This resulted in the detection of 39 misspecifications. Given the total number of constraints in the model – 1634 (86 constraints *19 countries) – 39 misspecifications is only a small percentage (2%). Only by chance alone one can expect a small percentage of misspecifications. The exact percentage is however difficult to give because some misspecifications are tied to each other, i.e. they represent equivalent models. In other words, not all misspecifications are independent (Saris, 2009). Nevertheless, we should judge every misspecification, but that does not mean we have to solve all misspecifications to have an acceptable model.

For Belgium, Germany, and Sweden we estimated a direct effect from the predictors to the item specific factors (s) to solve some misspecifications (see footnotes in table 5 for details). We also included a correlation between the unique components (u) of TRUST and FAIR in Israel and Portugal. After these re-specifications the model was estimated again and 24 misspecifications remained. In our view they were not serious enough and we ignored them in the further analysis.

The important model estimates are presented in table 5. The 5th and the 6th column contain the unstandardized estimates of the consistency coefficients. One can see that there is quite some variation across countries, which might lead to problematic issues when they are assumed equal for the metric invariance test. The average size of the consistency of Fair is 1.02 and for Help it is 0.79, for Trust it is 1 because that was the indicator used to define the latent scale of GST.

Table 5: Results of the configural invariance test[a].

		Consistency coefficients		
Group	**Country**	**Trust**	**Fair**	**Help**
1[b]	*Sweden*	1.00	1.04	0.82
2	*Austria*	1.00	1.11	0.92
3[b]	*Belgium*	1.00	0.95	0.79
4[b]	*Switzerland*	1.00	0.96	0.68
5	*Czech Republic*	1.00	0.95	0.77
6[b]	*Germany*	1.00	1.19	0.90
7	*Denmark*	1.00	0.91	0.71
8	*Spain*	1.00	0.91	0.71
9	*Finland*	1.00	0.98	0.86
10	*Great Britain*	1.00	1.09	0.87
11	*Greece*	1.00	1.03	0.85
12	*Ireland*	1.00	1.09	0.82
13[c]	*Israel*	1.00	1.02	0.70
14	*Italy*	1.00	1.14	0.90
15	*Netherlands*	1.00	0.90	0.68
16	*Norway*	1.00	0.98	0.67
17	*Poland*	1.00	1.05	0.80
18[c]	*Portugal*	1.00	0.99	0.62
19	*Slovenia*	1.00	1.09	0.87

[a] The presented figures are the unstandardized estimates. All estimates are significant at alpha=0.05. The consistency of TRUST is fixed to 1 to define the latent scale of GST.

[b] In these countries we estimated an extra direct effect: Education on TRUST (Belgium, Sweden), FeelSafe on TRUST (Germany).

[c] In these countries we estimated an extra correlated unique component: HELP with TRUST.

3.5.2. Results for metric invariance

In agreement with the suggested strategy (see section on *Strategy for measurement invariance testing*) we have selected Sweden as the reference country. The metric test involves the constraint as defined in equation 9. In this test the means play no role, so again all observed variables are expressed as deviation scores. All metric invariance constraints are tested with a delta2 of 0.15 and high power is 0.8 or larger. This resulted in the detection of 2 misspecifications for the metric invariance constraints. For the consistency of TRUST there is a misspecification in Norway (Cntry=16), but it is not possible to estimate this parameter because it is the reference indicator. In principle one should select another indicator as the reference indicator (Steenkamp & Baumgartner, 1998), however, that would only make sense if there is an indicator which is fully metric invariant; that's not the case. For the consistency of HELP there is a misspecification in Portugal (Cntry=18). This misspecification was solved by estimating that consistency coefficient not constrained to other consistency coefficients.

3.5.3. Result for scalar invariance

The test involves the constraint as defined in equation 10. In this test we also test the mean-structure, therefore the means of the observed variables are added to the model. The default procedure in the estimation of latent means is to fix the mean of the latent variables in the reference country to zero so that the latent means in the other countries are estimated relative to zero. We have, however chosen to fix the latent mean of GST in the reference country (Sweden) to the weighted mean of the indicators so that the estimated latent means can be more easily related to the scale on which the variables are measured. Finally, all scalar invariance constraints are tested with a delta4 of 0.07 and high power is 0.8 or larger.

The test for metric invariance resulted in one country, Portugal (Cntry=18), with a non-invariant consistency coefficient. In agreement with our own suggestion (see section on *Strategy for measurement invariance testing*) we have excluded, for Portugal, the intercept of HELP from the equality constraints. We estimated the model for all 19 countries and then analyzed the output with JRule, resulting in the detection of 3 misspecifications. Two for the intercept of TRUST in Belgium and Germany, and one for the intercept of Help in Ireland. The misspecification in Germany was rather large and we released the constraint on the intercept of TRUST in Germany. This resulted in a model with 2 misspecifications, i.e. in Belgium and Ireland. However, solving these misspecification and re-estimating the model again did not lead to changes in the other parameters, so we choose to accept those misspecifications.

3.6. Conclusion

The analysis of measurement invariance of GST indicates that the instrument available in the ESS is both partial metric and partial scalar invariant. This is very good news for studies - see references in the first sentence of this chapter - which assumed that the ESS measure of GST is comparable. In those studies however, the variable GST was created as composite score, and for composite scores it is imperative that GST shows full metric and scalar invariance (Saris and Gallhofer, 2007, ch. 16). That is not the case here, but the number of non-invariant parameters is so small – 3 in total – that it is unlikely to have a significant effect on the comparability if composite scores are used. It is however easier to continue analyzing the data in the framework of structural equation modeling, treating GST as a latent variable. Because in that framework it is possible to correct for misspecifications, e.g. non-invariant parameters, so that these misspecifications do not bias parameter estimates. Note that this is not the same as saying that the parameter estimates are unbiased after misspecifications are solved, that can only be true if other model assumptions, e.g. error-free observations, are not violated.

Now that we established that the measure of GST is partial metric and partial scalar invariant we can make cross-national comparisons in two ways. We can make a ranking of the level of GST, the latent variable, in the 19 countries. We can also study, in principle, the causes of GST cross-nationally. Table 6 ranks the countries in our analysis on their level of GST. The results are more or less in line with earlier studies on data from the World Values Studies (WVS) by Van Deth (2001), Norris (2002), and Inglehart (1999), on data of the European Values Studies (EVS) by Adam (2008), on data from the Euromodule survey by by Delhey & Newton (2005), and on data from the ESS by Reeskens & Hooghe (2008) and Zmerli & Newton (2008). Note however, that studies based on the WVS only use a single question to measure GST and therefore produce very different levels of trust. Nevertheless, the rank-order of the countries is rather similar.

Table 6: The estimates of the latent means of GST[a].

Country	**Mean GST**
Denmark	6.81[b]
Norway	6.51[b]
Finland	6.34[b]
Sweden	5.97[b]
Switzerland	5.71[b]
Netherlands	5.70[b]
Ireland	5.69[b]
Germany	5.30[b]
Great Britain	5.27[b]
Austria	5.27[b]
Belgium	5.02[b]
Israel	4.91[b]
Spain	4.83[b]
Portugal	4.60[b]
Czech Republic	4.49[b]
Italy	4.36[b]
Slovenia	4.29[b]
Poland	3.83[b]
Greece	3.43[b]

[a] Countries are sorted in descending order of the means.

[b] Significant at $\alpha=0.05$

While ranking the countries on their mean levels could be an interesting exercise to describe where countries are, it does not explain much. Our initial interest was more in exploring the causes (and level) of GST. Previous studies have resulted in mixed conclusions, as described in detail by Newton (2004). He mentions that there is no single theory which holds across most countries, i.e. some theories work in some countries but not in other. It is our belief that such conclusions are not warranted, because the predictors in those studies contained, most certainly, measurement error. This is also the reason why we dare not draw any conclusion from our model in this respect. Despite that, we have presented the results in table 7

Table 7: The completely standardized total effect of the predictors on GST[a].

Country	**Happy**	**FlSafe**	**MeetEnd**	**Social**	**Religs**	**Urban**	**Educ**	**Age**	**Gndr**	**R^2**
Sweden	.21[b]	.21[b]	-.09[b]	.08[b]	.08[b]	.01	.08[b]	.12[b]	.16[b]	16%
Austria	.22[b]	.13[b]	-.11[b]	.03	.02	-.02	.07[b]	-.04[b]	.03	13%
Belgium	.16[b]	.12[b]	-.19[b]	.10[b]	.08[b]	-.03	.15[b]	.02	.02	17%
Switzerland	.24[b]	.16[b]	-.13[b]	.11[b]	.03	-.01	.13[b]	.13[b]	.05	16%
Czech Rep.	.21[b]	.15[b]	-.04	.05	.02	.02	.15[b]	.03	.05[b]	12%
Germany	.23[b]	.09[b]	-.11[b]	.10[b]	.06	.03	.07[b]	.09[b]	.06[b]	13%
Denmark	.21[b]	.12[b]	-.08[b]	.02	.02	.04	.20[b]	.10[b]	.18[b]	15%
Spain	.14[b]	.14[b]	-.12[b]	.01	.01	.03	.12[b]	.05	.00	8%
Finland	.26[b]	.19[b]	-.13[b]	.06	.10[b]	.08[b]	.03	.06	.09[b]	17%
Great Britain	.21[b]	.19[b]	-.03	.05	.07[b]	-.03	.11[b]	.19[b]	.03	15%
Greece	.18[b]	.14[b]	-.02	.02	-.07[b]	.04	.09[b]	.02	.05[b]	9%
Ireland	.21[b]	.12[b]	-.07[b]	.06[b]	.03	.00	.12[b]	.15[b]	.04	12%
Israel	.21[b]	.03	-.10[b]	.12[b]	.03	-.03	.15[b]	.13[b]	.07[b]	13%
Italy	.18[b]	.13[b]	-.12[b]	.07[b]	.00	-.04[b]	.15[b]	.07[b]	.07[b]	14%
Netherlands	.17[b]	.15[b]	-.09[b]	.03	.04	-.06[b]	.15[b]	.04	.08[b]	12%
Norway	.23[b]	.16[b]	-.06[b]	.12[b]	-.04	-.05	.17[b]	.20[b]	.20[b]	18%
Poland	.27[b]	.11[b]	-.04	.10[b]	.08[b]	.10[b]	.11[b]	-.01	.07[b]	16%
Portugal	.17[b]	.19[b]	.05	.11[b]	-.06[b]	-.07[b]	.15[b]	.18[b]	.04	11%
Slovenia	.21[b]	.08[b]	-.09[b]	.08[b]	.04	-.02[b]	.12[b]	.08[b]	.00	11%

[a] *All estimates are within group completely standardized estimates.*

[b] *Significant at α=0.05.*

Table 7 holds the figures that provide an answer to which (individual level) theories explain GST. Even though we have our methodological reservations, we will interpret some of the results, but conclusion should be taken with a *grain of salt*. The standardized estimates in table 7 reveal that the effect of each predictor on GST is pretty consistent across the countries. The variable Happy (success and well-being theory) explains most of GST. Then there are several variables that explain GST a bit less well (in order of importance): the subjective experience of the neighborhood (FlSafe), whether people find it difficult to make ends meet (MeetEnd), and the frequency of social contacts (Social). The following two variables do not contribute at all to GST the size of the community (Urban) and Religiosity

(Religs). The effect of Gndr is at average very small, with three eye-catching exceptions, Sweden, Norway and Denmark. One would like to speculate – but we don't – why these countries are so deviant, because they are geographically and politically close to each other. Finally, the variable education explains at average 1.6% of the variance in GST, which makes this a relatively important variable in comparison with the others. This finding is in contrast to what Delhey & Newton (2003) found, who reported to their own surprise no effect of education. A very tentative overall conclusion would be that in contrast to Delhey & Newton (2003), we find rather consistent cross-national results. That is, Happy is always the most important cause. In addition, the standardized effect of all predictors are more or less the same cross-nationally. However, these conclusions are very tentative for the reasons we mentioned earlier.

4. Discussion

The reader might have the feeling that this chapter does not live up to the expectation set at the start of this chapter. We aimed at carrying out a cross-national analysis of the causes of GST while also introducing two innovations, i.e. model testing and correction for random and systematic measurement error in measurement invariance tests. The disappointment lays – for some part – in the fact that in the end we did not correct for systematic measurement error in the indicators of GST and we also did not draw *real* conclusions from the causal analysis. Hence, did we choose the wrong topic to illustrate our innovations? Certainly not! Generalized social trust is believed to be at the heart of a healthy society (Uslaner, 2000; Uslaner, 1999; Putnam, 1993) which justifies our choice. The fact that we did not correct for systematic measurement error in the end was related to the fact that for all but 3 countries the method factor variance, due to the measurement procedure, was not significant. This result was expected, because multitrait multimethod experiments in the ESS pilot study indicated that the response scales and formulation of the items to measure GST produced little if any method factor variance (Saris & Gallhofer, 2003). That was also the reason they were included in the main ESS questionnaire. In addition, the – more or less – absence of systematic measurement error allowed us to simplify the presentation and discussion of the models. Another part of the disappointment is grounded in our reluctance to seriously interpret the results of the causal analysis. The reason for that is simply that an important assumption of the model, i.e. exogenous variables are observed without measurement error, was violated. This assumption is quite often ignored in research. The negative consequences of this neglect will be illustrated with a simple example.

Figure 5: Regression estimates when x1 and x2 are assumed to be observed without measurement error.

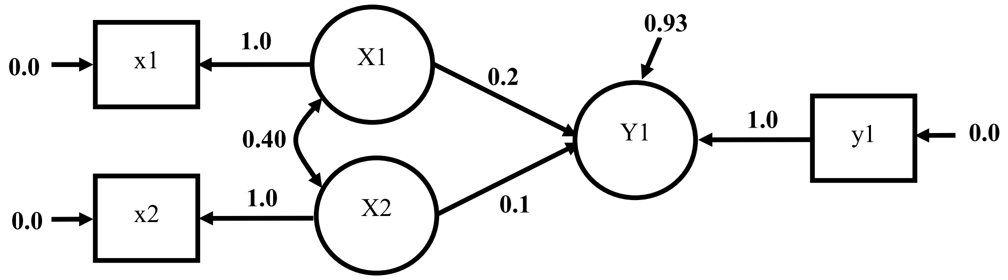
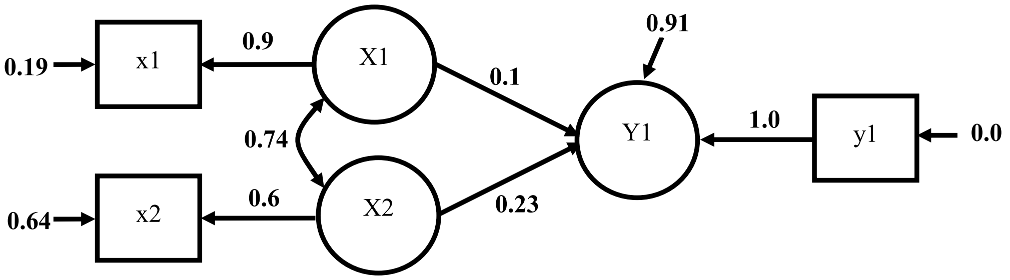


Figure 6: Regression estimates taking into account that x1 and x2 are observed with measurement error. Reliability of x1 is 0.81 and of x2 is 0.36, i.e. the square of the factor loadings.



In the figures above a regression model is presented with three observed variables (y1,x1,x2). The observed correlations between these variables are: 0.24, 0.18 and 0.40. For the sake of simplicity we let y1 be observed without errors, but x1 and x2 both contain random measurement error. The regression estimates, ignoring the presence of measurement error, are presented in figure 5. One can see that x1 is the most important predictor with a standardized effect of 0.20 on y1. However, due to the presence of measurement error things can change dramatically. Let the measure of x1 have a reliability of 0.81, and the measure of x2 a reliability of 0.36. The latter reliability is not particularly good, but it is also not uncommon to have indicators with a loading of 0.6, which actually is a reliability of 0.36. Figure 6 presents the same model as before, but now corrected for measurement error. It is immediately clear that the substantive conclusions are very different. It is now variable x2 which is the most important predictor with an effect of 0.23. Because we realize that something similar could occur in our analysis of GST, we did not dare to draw substantive conclusions. Nevertheless, that was very tempting given that the results are in line with our expectations that *the same causal theories are at work across countries and that each theory has approximately the same importance across countries*.

In this chapter we concluded that the instrument to measure GST, as used in the ESS, is partial scalar invariant. This means that we can compare both means and relationships cross-nationally of this measure with other scalar invariant measures. This is very good news for studies that assumed that GST is comparable across the ESS. Reeskens & Hooghe (2008) arrived, however, at a different conclusion. That is, they seriously questioned the scalar invariance of this instrument. This is strange because their conclusion is based on the same data. The answer to why the conclusions are so different are related to our innovations. We correct for random (and systematic) measurement error in the indicators of the instrument, they don't. Hence, their metric and scalar invariance tests are concerned with different parameters. We evaluate our models through the detection of misspecifications, taking into account the power of the test, they evaluate their models with the RMSEA, the NNFI, and the modification index (MI). We have illustrated in Saris et al. (2009) that the RMSEA and the MI are sensitive to the power. When the power is very high, the RMSEA and the MI tend to over-reject models. We have seen in our analysis (unreported findings of this study) that the inclusion of a mean structure in the model increases the power of the test considerably, which can explain why Reeskens & Hooghe (2008) were forced to reject the scalar invariance of GST.

Finally, we should say something about the alternative test procedure because it is closely linked to our successful attempt to illustrate the partial scalar invariance of GST. Our procedure to evaluate structural equation models has two distinct features. First of all, the procedure takes the power of the test into account. Second, the procedure allows – in a sense - for non-exact testing, i.e. the test ignores – in principle – misspecifications that are smaller than delta. This way, a distinction can be made between relevant and irrelevant misspecifications. What relevant and irrelevant is, is not easy to define but by and large one could say that misspecifications which do not alter the substantive conclusions are irrelevant.

As a rule of thumb 0.10 is considered a relevant misspecification for correlations and effects, and 0.40 for factor loadings (Saris et al., 2009). For the correlations between random error terms (e) we consider a misspecification of 0.10 as relevant, although, given the random nature of these errors we believe such misspecifications cannot be present. For correlations between unique components (u) we consider a misspecification of 0.15 as relevant. The reason that we allow for a larger misspecification in this case (compared to the rule of thumb-value 0.10) is that we correct for the presence of measurement error. If we correct a correlation of 0.10 for attenuation using the average the quality of single item measure, which is 0.8 (Saris & Gallhofer, 2007; Scherpenzeel, 1995b), we obtain a corrected correlation of 0.15 (=0.10/0.8*0.8). We have used the same value (0.15) for regression effects from the predictors (x) on the item specific factors (s). For the metric and scalar invariance tests we test for the equality factor loadings and on intercepts. For such equality constraints there are no rules of thumb. In spite of a lack of rules, we have come up with a set of values for delta that we believe are relevant misspecifications. In order to determine what a relevant misspecification for the equality constraints on the factor loadings and intercepts is, we used the results from the study of De Beuckelaer & Swinnen (this book). They found that the probability of drawing an incorrect conclusion that two countries have different (or the same) latent means increases strongly under the following two conditions. First, if factor loadings deviate more than 30% from the population value and second if intercepts deviate more than 10% of the length of the response scale from the population value. Because their analysis was on standardized variables, these percentages correspond to relevant misspecifications of 0.30 and 0.10. However, in order to be on the safe side, we decided to test for misspecifications that were only half the magnitude that would follow from the study of De Beuckelaer & Swinnen (this book). Hence, we considered 0.15 (or larger) a relevant misspecification for factor loadings that are constrained to equality, and 0.07 (or larger) a misspecification for intercepts that are restricted to be the same. Please note that the values for relevant misspecifications (deltas) are standardized values, but normally, we analyze unstandardized variables. These standardized deltas will, however, be unstandardized in JRule using the scales and variances of the variables in the model.

Notes:

[1] It is possible to estimate the intercepts τ_s if the mean of the latent variable (f) is fixed to zero as well as the intercepts τ_y of the indicators, but in that case the intercepts will be equal to the observed means. In that case we could search for the intercept which is most deviant from the estimate, nevertheless, the procedure suggested in the text does exactly that, i.e. inspection of the residuals of the means.

[2] JRule is currently freeware and can be obtained by sending a request to the first author of this chapter via e-mail (w.vanderveld@socsci.ru.nl).

[3] It should be possible to estimate the latent means of the common method factors in the context of measurement invariance testing, using an MTMM or split ballot MTMM design. However, the models that have to be specified will be very complex, making this an unattractive solution.

[4] Most missing data were missing by design, because of the split ballot nature of the MTMM experiments. Only a small percentage of the data were not missing by design and we assumed they were missing at random.

REFERENCES

Adam, Frane (2008). Mapping social capital across Europe: findings, trends and methodological shortcomings of cross-national surveys. *Social Science Information* 47(2): 159-186.

Andrews, Frank M. (1984). Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach. *Public Opinion Quarterly* 48(2):409-42.

Barrett, Paul (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences* 42(5):815–824.

Batista-Foguet, Joan M., and Coenders, Germà (2000). Modelos de Ecuaciones Estructurales. Madrid: La Muralla.

Blalock, Hubert M. Jr. (1968). The measurement problem: A gap between languages of theory and research. In: Blalock Hubert M., and Blalock, Ann B. eds., *Methodology in the Social Sciences*. London: Sage.

Brown, Michael W., and Cudeck, Robert (1993). Alternative ways of assessing model fit, Pp. 136-162. In: Bollen, Ken, Long, J. Scott, eds., *Testing Structural Equation Models*. London: Sage.

Byrne, Barbara M., Shavelson, Robert J., and Muthén, Bengt (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin* 105(3):456-466.

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

Corten, Irmgard W., Saris, Willem E., Coenders, Germà, Van der Veld, William M., Aalberts, Chris E., and Kornelis, Charles (2002). Fit of Different Models for Multitrait–Multimethod Experiments. *Structural Equation Modeling: A Multidisciplinary Journal* 9(2):213–232.

Delhey, Jan, and Newton, Kenneth (2005). Predicting Cross-National Levels of Social Trust: Global Pattern or Nordic Exceptionalism? *European Sociological Review* 21(4):311-327.

Delhey, Jan, and Newton, Kenneth (2003). Who trusts? The causes of social trust in seven societies. *European Societies*, 5(2):93-137.

Enders, Craig K., and Bandalos, Deborah L. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal* 8(3):430-457.

Erikson, Erik H. (1950). *Childhood and Society*. New York: Norton.

Fukuyama, Francis (1995). *Trust: The Social Virtues and the Creation of Prosperity*. New York: The Free Press.

Ganzeboom, Harry B. G. (2009). Multiple Indicator Measurement of Social Background. *Keynote presented at the 3rd conference of the European Survey Research Association*, Warsaw, Poland.

Gesthuizen, Maurice, Scheepers, Peer, Van der Veld, William M., and Völker, Beate (in press). Structural aspects of social capital: Tests for cross-national equivalence in Europe. *Quality and Quantity*.

Heise, D. R., & Bohrnstedt, G. W. (1970). Validity, Invalidity, and Reliability. *Sociological Methodology* 2:104-129.

Herreros, Francisco, and Criado, Henar (2008). The State and the Development of Social Trust. *International Political Science Review* 29(1):53–71.

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

Inglehart, Ronald (1999). Trust, Well-being, and Democracy, Pp. 88-121. In: Warren, Mark E., *Democracy and Trust*. Cambridge University Press.

Johnson, Emily C., Meade, Adam W. and DuVernet, Amy M. (2009) The Role of Referent Indicators in Tests of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal* 16(4):642-657.

Jöreskog, Karl G., and Sörbom, Dag (1996). *LISREL® 8 User's reference guide*. Chicago, Illinois: Scientific Software International.

Kaasa, Anneli, and Parts, Eve (2008). Individual-Level Determinants of Social Capital in Europe. Differences between Country Groups. *Acta Sociologica* 51(2): 145–168.

Kaase, Max (1999). Interpersonal trust, political trust and non-institutionalised political participation in Western Europe. *West European Politics* 22(3):1-21.

Kaplan, David (1989). Model Modification in Covariance Structure Analysis: Application of the Expected Parameter Change Statistic. *Multivariate Behavioral Research 24* (3):285-305.

Knack, Stephen, and Keefer, Philip (1997). Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *The Quarterly Journal of Economics* 112(4):1251-1288.

Kumata, Hideya, and Schramm, Wilbur (1956). A Pilot Study of Cross-Cultural Meaning. *Public Opinion Quarterly*, 20(1):229-238.

Letki, Natalia, and Evans, Geoffrey (2008). Endogenizing Social Trust: Democratization in

East-Central Europe. *British Journal of Political Science* 35:515–529.

Levi, Margaret (1996). Social and Unsocial Capital: A Review Essay of Robert Putnam's Making Democracy Work. *Politics and Society* 24(1):45-55.

MacCallum, Robert C., Browne, Michael W., and Sugawara, Hazuki M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1(2): 130-149.

Marsh, Herbert W., Hau, Kit-Tai, and Wen, Zhonglin (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal* 11(3):320-341.

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

Millsap, Roger, E., and Meredith, William (2007). Factorial Invariance: Historical Perspectives and New Problems, Pp 130-152. In: Cudeck, Robert, and MacCallum, Robert C. eds., *Factor Analysis at 100: Historical Developments and Future Directions*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc. Publishers.

Muthén, Linda K., and Muthén, Bengt O. (1998-2007). *Mplus User's Guide. Fifth Edition.* Los Angeles, CA: Muthén & Muthén

Newton, Kenneth (2004). Social Trust: Individual and Cross-National Approaches. *Portuguese Journal of Social Science* 3(1):15-35.

Newton, Kenneth (1997). Social capital and democracy. *American Behavioral Scientist* 40(5):575–86.

Norris, Pipa (2002). Making Democracies Work: Social Capital and Civic Engagement in 47 Societies. Paper presented at the *Midwest Political Science Association 60th Annual Meeting*, Chicago, United States.

Northrop, Filmer S.C. (1969). *The Logic of the Sciences and the Humanities*. Cleveland: World Publishing Company.

Paxton, Pamela (2002). Social Capital and Democracy: An Interdependent Relationship. American Sociological Review 67(2):254-277.

Putnam, Robert D. (2000). *Bowling alone: The collapse and revival of American community*. New York: Simon & Schuster.

Putnam, Robert D. (1993). *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton, New Jersey: Princeton University Press.

Reeskens, Tim, and Hooghe, Marc (2008). Cross-cultural measurement equivalence of generalized trust. Evidence from the European Social Survey (2002 and 2004). *Social Indicators Research* 85(3):515-532.

Rensvold, Roger B., and Cheung, Gordon W. (2001). Testing for Metric Invariance using Structural Equation Models. Solving the Standardization Problem, Pp. 25-50. In: Schriesheim, Chester A., and Neider, Linda L. eds., *Equivalence in Measurement*. Greenwich, Connecticut: Information Age Publishing.

Rothstein, Bo, and Uslaner, Eric M. (2005). All for All: Equality and Social Trust. *LSE Health and Social Care Discussion Paper No. 15*. SSRN: http://ssrn.com/abstract=824506.

Saris, Willem E. (2009). Some Important Considerations while Testing for Misspecifications in SEM. *Study presented at the 3rd conference of the European Survey Research Association*, Warsaw, Poland.

Saris, Willem E., and Andrews, Frank M. (1991). Evaluation of measurement instruments using a structural modeling approach, Pp. 575-599. In: Biemer, Paul P., Groves, Robert M., Lyberg Lars E., et al. Eds., *Measurement Errors in Surveys*. New York: John Wiley & Sons.

Saris, Willem E., and Gallhofer, Irmgard N. (2003). Report on the MTMM Experiments in the Pilot Studies and Proposals for Round 1 of the ESS. Made for the Central Coordinating Team of the European Social Survey, London.

Saris, Willem E., and Gallhofer, Irmtraud N. (2007). *Design, Evaluation and Analysis of Questionnaires for Survey Research*. Hoboken, New Jersey: John Wiley & Sons.

Saris, Willem E., Revilla, Melanie, Krosnick, Jon A., Shaeffer, Eric M, (in press). Comparing Questions with Agree/Disagree Response Options to Questions with Construct-Specific Response Options. *Public Opinion Quarterly*, forthcoming.

Saris, Willem E., Satorra, Albert, and Sörbom, Dag (1987). The detection and correction of specifications errors in structural equation models. *Sociological Methodology*. 17:105-129.

Saris, Willem E., Satorra, Albert, and Coenders, Germà (2004). A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design. *Sociological Methodology* 34:311-347.

Saris, Willem E., Satorra, Albert and Van der Veld, William M. (2009). Testing Structural Equation Models or Detection of Misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal* 16(4): 561-582.

Saris Willem E., Van der Veld, William M., and Gallhofer, Irmtraud N. (2004). Development and Improvement of Questionnaires Using Predictions of Reliability and Validity, Pp. 275-298. In Presser, Stanley, Rothgeb, Jennifer M., Couper, Mick P., Lessler, Judith T., Martin, Elizabeth, Martin, Jean, and Singer, Eleanor, eds., *Methods for Testing and Evaluating Survey Questionnaires*. New York: John Wiley & Sons.

Satorra, Albert, and Bentler, Peter M. (1994). Corrections to test statistics and standard errors in covariance structure analysis, Pp. 399-419. In: Von Eye, Alexander & Clogg, Clifford eds., *Latent Variables Analysis: Applications for Developmental Research*. Thousand Oaks, California: Sage.

Scherpenzeel, Annette C. (1995a). *A Question of Quality. Evaluating Survey Questions by Multitrait-Multimethod Studies*. Dissertation. Koninklijke PTT Nederland NV, KPN Research.

Scherpenzeel, Annette C. (1995b). Life satisfaction in The Netherlands. In: Saris, Willem E., Veenhoven, Ruut, Scherpenzeel, Annette C., and Bunting, Brendan, eds., *A comparative study of satisfaction with life in Europe*. Budapest: Eötvös University Press.

Scherpenzeel, Annette C., and Saris, Willem E. (1997). The validity and reliability of survey questions: a meta-analysis of MTMM studies. *Sociological Methods and Research* 25(3):341-383.

Steenkamp, Jan-Benedict E.M., and Baumgartner, Hans (1998). Assessing measurement equivalence in cross-national consumer research. *Journal of Consumer Research* 25(1):78-90.

Steyer, Rolf, and Schmitt, Manfred J. (1990). Latent state-trait models in attitude research. *Qality and Quantity* 24(4):427-445.

Torcal, Mariano, and Montero, José R. (1999). Facets of social capital in new democracies, Pp. 167-191. In: Van Deth, Jan, Maraffi, Marco, Newton, Kenneth, and Whiteley, Paul, eds., *Social Capital and European Democracy*. London: Routledge.

Uslaner, Eric M. (1999). Democracy and Social Capital, Pp. 121-150. In: Warren, Mark E. ed., *Democracy and Trust*. Cambridge University Press.

Uslaner, Eric M. (2000). Producing and Consuming Trust. *Political Science Quarterly* 115(4):569-590.

Van Deth, Jan W. (2001). The Proof of the Pudding: Social Capital, Democracy, and Citizenship. *Paper prepared for delivery at the EURESCO Conference on 'Social Capital: Interdisciplinary Perspectives*'. Exeter, United Kingdom.

Van der Veld, William M. (2006). *The Survey Response Dissected. A new theory about the survey response process*. Dissertation, University of Amsterdam: Amsterdam School for Communications Research (ASCoR).

Van der Veld, William M., and Saris, Willem E. (2004). Separation of reliability, validity, stability, and opinion crystallization. In: Saris, Willem E., and Sniderman, Paul M. eds., *Studies in Public Opinion: Attitudes, Nonattitudes, Measurement Error, and Change*. Princeton, New Jersey: Princeton University Press.

Van der Veld, William M., Saris, Willem E., and Satorra, Albert (2008). JRule 3.0: User's Guide. http://www.vanderveld.nl/JRule.

Yoon, Myeongsun, & Millsap, Roger E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal* 14(3):435-463.

Zmerli, Sonja, and Newton, Kenneth (2008). Social Trust and attitudes toward democracy. *Public Opinion Quarterly* 72(4):706-724.