Developments in Social Science Methodology
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 18, Ljubljana: FDV, 2002

Equivalence of Measurement Instruments for Attitude Variables in Comparative Surveys, Taking Method Effects into Account: The Case of Ethnocentrism

Jaak Billiet[1], Bart Cambré[2], and Jerry Welkenhuysen-Gybels[3]

Abstract

This study is focused on the construction of a cross-national comparable measurement instrument for attitude variables in comparative surveys. Multi-group measurement models for latent variables (LISREL), taking method effects into account, are applied. The measurements of the 'out-group' dimension of ethnocentrism (variables q42, q44, q45, q47-q52) in the 1995 ISSP dataset are used. Nearly all the items in the quasi balanced set are written in a Likert format in which respondents are asked how strongly they agree or disagree with each attitude statement. There is considerable evidence that such a response format can be susceptible to an agreeing-response bias called acquiescence (Billiet and McClendon, 2000). It is shown that in all countries, models with a method or style factor (acquiescence) always fit the data better than models without a style factor. It is investigated to what extent the measurement instrument with a content and a method factor is equivalent over the cultural groups. In a first step the factor loadings of the groups are explored by cluster analysis. After the detection of two subsets of groups that are likely to share equivalent measurement instruments, a stepwise procedure was performed starting with the measurement model for one group, and then looking for equivalent groups (countries) only accepting minor changes in the measurement model. The introduction of a style factor allows us to control for a possible source of measurement non-equivalence, namely method bias. Moreover, the inclusion of a method factor gives the opportunity to investigate the differences in method effects between the groups (countries). However, it is found that both, the variance of the style factor and its factor loadings do not differ between the groups in the first subset of countries, but there are

[1] Department of Sociology, Katholieke Universiteit Leuven, E. Van Evenstraat 2B, 3000, Leuven, Belgium.

[2] Centre for Data Collection and Analysis, Department of Sociology, Katholieke Universiteit Leuven.

[3] Fund for Scientific Research (Flanders), Centre for Data Collection and Analysis, Department of Sociology, Katholieke Universiteit Leuven.

differences in the degree of acquiescence in the second one. It seems reasonable to conclude that the agreeing-response bias does not lead to a cross-cultural method bias in the measurement of ethnocentrism in the Western countries of ISSP 1995.

1 Introduction

As sociological survey research is becoming more and more concerned with the comparison of concepts over different cultures and/or nations, the importance of obtaining adequate measures for those concepts in each of these groups has become clear. Besides the methodological problems that already occur when doing intra-cultural research, the researcher interested in cross-cultural comparisons also has to deal with specific problems which are inherent to the latter kind of research. One cannot readily assume that the scores of respondents of different cultural groups on a certain item or a certain scale, can be compared in a direct and straightforward way. The comparability of scores obtained in such a manner depends on their level of equivalence.

In literature, several definitions of equivalence can be found. Johnson (1998) classifies the definitions of this term into two categories. *Interpretative equivalence* deals with similarities in the way abstract, latent concepts are interpreted among different cultures or cultural groups. One of the most cited forms of interpretative equivalence is called concept equivalence (Hui and Triandis, 1985). This kind of equivalence implies that the concept can be meaningfully discussed in the cultures or cultural groups concerned. Concept equivalence is, of course, a prerequisite for every cross-cultural comparison. With the term *procedural equivalence*, Johnson (1998) refers to the types of equivalence that are concerned with the measurements and procedures used to make cross-cultural comparisons. In their definition of procedural equivalence, Berry et al. (1993: 238) made a distinction between the *comparison scale* and the *measurement scale* of a theoretical concept. The comparison scale is a hypothetical scale that is postulated for the concept of interest. Because of the hypothetical nature of this scale, a measurement scale is necessary to measure the concept. Non-equivalence, then, implies that observed differences on the measurement scale between the cultural groups don’t correspond to differences on the comparison scale.

Van de Vijver and Leung (1997: 8-9 ; van de Vijver, 1998) distinguish three levels of procedural equivalence which are hierarchically related to each other. The first and lowest level of equivalence is called *construct equivalence* and is achieved if the instrument measures the same latent trait in all of the cultural groups under investigation. *Measurement unit equivalence* is the next and higher level of equivalence. This level of equivalence is obtained if the measurement unit of the instrument is identical for each of the cultural groups (van de Vijver and Leung, 1997 ; van de Vijver, 1998). If the measurement scales of the cultural groups also have the same origin then the highest level of equivalence, *scalar equivalence* or *full score comparability*, is attained. Only this last level of equivalence permits the researcher to directly compare the scores of different cultural groups. However there are three types of bias that can influence the comparability of scores obtained across cultural groups (van de Vijver and Leung, 1997; van de Vijver, 1998). *Construct bias* is characterised by dissimilarities in the operationalization of the concept across cultures. *Method bias* can either arise from the incomparability of the samples, instrument characteristics to which individuals from different cultures react in a consistently different manner or differences in the administration of the instrument. The third kind of bias, *item bias* or *differential item functioning*, is caused by anomalies at the item level such as poor translation, incidental differences in the response scale, etc.

The scope of this article is to evaluate the equivalence of the construct that was used to operationalize feelings of threat towards immigrants across 9 European countries. Several methods exist for evaluating the construct equivalence of a certain construct (van de Vijver and Leung, 1997; Welkenhuysen-Gybels, 1998). Here, structural equation modelling is used because it also allows us to control for a specific type of method bias, called acquiescence. After all, the items of the ethnocentrism scale have identical Likert-type response scales in which respondents are asked how strongly they agree or disagree with each attitude statement. Several sociologists have argued that there is considerable evidence that such a response format can be susceptible to an agreeing-response bias called acquiescence. This agreeing-response bias can be defined as the tendency to agree with statements or questions, independent of their content. Billiet and McClendon (1998, 1999) have shown that it is possible to use structural equation modelling to control for this type of response effect. They specified an extra style factor (or common method covariance) which substitutes a number of previously unidentified error covariances. The identification of this response effect, however, requires that the attitude scale is balanced. This means that it should contain a more or less equal number of positively and negatively worded items. Respondents with a tendency towards “yeah-saying” will agree with the positively as well as the negatively worded items.

Therefore, it will first be evaluated whether such an agreeing-response bias can be discerned in the countries concerned. This implies that we will first check whether there is an intra-cultural method effect; whether the responses to the 9 items under investigation are susceptible to the agreeing-response bias discussed above. If this is the case, we can also evaluate whether there is a method bias in the cross-cultural sense; in other words: whether the method effect differs across countries. After all, it is very possible that respondents from different cultures differ in their susceptibility to an agreeing-response bias. Either way, controlling for acquiescence will lead to a more valid assessment of the equivalence of the construct of interest.

2 Methods

The analyses will be performed on (part of) the data set of the 1995 International Social Survey Program (ISSP).[4] However, not all 23 ‘countries’ in which the

[4] The researcher wish to thank the researchers of ISSP for the delivery of the 1995 data.

survey was conducted, will be taken up in the analysis. Only the Western European countries will be taken into account, namely: Austria, Ireland, Italy, The Netherlands, Norway, Spain, Sweden, The United Kingdom and (former) West Germany[5].

Table 1: Question wordings of the selected questions (British version of the questionnaire).

Item		Question wordings
Q42	(-)	Foreigners should not be allowed to buy land in [*country*].
Q44	(-)	It is impossible for people who do not share [*country's*] customs and traditions to become fully [*countryman*]
Q45	(+)	Ethnic minorities should be given government assistance to preserve their customs and traditions.
Q47	(-)	Immigrants increase crime rates
Q48	(+)	Immigrants are generally good for [*country's*] economy.
Q49	(-)	Immigrants take jobs away from people who live in [*country*]
Q50	(+)	Immigrants make [*country*] more open to new ideas and cultures.
Q51	(n)	Do you think the number of immigrants to [*country*] nowadays should be … (increased – reduced)
Q52	(+)	How much do you agree or disagree that refugees who have suffered political repression in their own country should be allowed to stay in [*country*]

The questionnaire contained (a translation of) a number of items in Likert format[6] about the respondent's feelings of threat towards immigrants, foreigners, or ethnic minorities. Only four of these are commonly conceived as indicators for the attitude toward immigrants (the statements Q47, Q48, Q49, and Q50 in the ISSP questionnaire). However, we expected that other items about ethnic minorities or foreigners that are placed in the same context of the questionnaire are measuring a global concept expressing a general attitude towards 'outgroups'.[7] Therefore we have chosen a larger balanced set of eight items. The English version of these items is shown in Table 1. As required for the detection of acquiescence, the scale is balanced: it contains as many positively (Q45, Q48, Q50, Q52) as negatively worded items (Q42, Q44, Q47,Q49). We also included

[5] The East German and the West German sample were not combined, because of the cultural-historical differences between these two countries.

[6] Likert-scales have response scales running from "strongly agree" to "agree", "uncertain", "disagree" and "strongly disagree"? These five positions are given simple weights of 5, 4, 3, 2, and 1. The "uncertain" category is often replaced by "neither agree nor disagree" (Likert, 1932).

[7] This expectation is based on our experience with sets of 14 items in the questionnaires of the 1991 and 1995 *Belgian General Election Surveys*. All items about foreigners or immigrants tend to form one global factor.

one item (Q51) which can be expected to be unaffected by the agreeing response bias, because it doesn't have a Likert-type format. For this item, the respondents had to indicate whether they thought the number of immigrants that was allowed into their country should be increased or decreased.

Table 2: Scheme of the imputation of missing values.

Country	Original number of observations	Remaining number of observations	Number of observations with 1 imputation	Number of observations with 2 imputations	Number of excluded observations
Austria	1,007	926	170	42	81
(West) Germany	1,282	1,110	237	50	172
Ireland	994	961	133	16	33
Italy	1,094	1,053	138	17	41
The Netherlands	2,089	1,000 (+908)*	255	55	181
Norway	1,527	1,372	243	62	155
Spain	1,230	1,076	197	40	154
Sweden	1,296	1,143	215	54	153
United Kingdom	1,058	979	145	26	79

* A random sub-sample of the Netherlands (N = 908) was used for exploration of the models.

To avoid loss of data due to missing values, an imputation technique was used. The imputation was conducted separately for the negatively and the positively worded items. For the positively worded items and the neutral item (Q51), missing values were only imputed if the respondent had one missing value on these five items. In this case the missing value was replaced by the mean value of the responses on the other four items. Respondents with more than one missing value were excluded from the analysis. An analogous way of working was used for the negatively worded items. Hence, at most 2 missing values were imputed per observation. A scheme of the results of the imputation can be found in Table 2.

From Table 2, one can see that the samples are about the same size, except for the Dutch sample, which is a lot larger than the other samples. This could lead to an over-determination of the model selection process by the covariance structure of the Dutch sample. After all, model improvements are based on the modification indices that are provided by LISREL (version 8.30 was used for our analyses). These modification indices show the virtual drop in the Chi-square statistic if the respective parameter would be estimated freely. It is, however, well known that the Chi-square statistic is very sensitive to the size of the sample. This means that, ceteris paribus, a larger sample will yield a larger value for the Chi-square statistic. To avoid this problem, a simple random sample of 1,000 units was drawn from the Dutch sample (after imputation). The remaining sub-sample of the Netherlands (N = 908) was used for the exploration of the two basis models (with and without a style factor).

The multigroup analyses to evaluate the construct equivalence of the substantive scale are based on the Pearson covariance matrix and the maximum likelihood estimation (ML) procedure. In the past, there has been a large discussion on the appropriateness of this procedure for Likert Type items (O'Brien, 1985; Homer and O'Brien, 1988). For this kind of items, some scholars propose another procedure, namely polychoric correlations with equal thresholds in each group.[8] The multigroup analyses should then be performed on the asymptotic covariance matrix and a weighted least squares estimation (WLS) procedure should be used (Jöreskog, 1990). This way of working was tried out for some of the multigroup analyses. The results were very similar to those of the first procedure, however the latter procedure resulted in larger number of error covariances and a much weaker fit of the models. Moreover, in the light of the evaluation of equivalence, it is not clear what the implications are of forcing the responses in different cultural groups a priori into the same underlying response scales. Therefore the first procedure was used (ML estimations on the Pearson covariance matrix). On the basis of simulations, several authors defend this procedure for ordinal scored data if one is interested in the latent variables and if the distributions of the observed indicators are not too skewed (Johnson and Creech; 1983; Coenders, 1996: Coenders and Saris, 1995).[9]

3 Improving the measurement model by including a style factor

In a recent study, Billiet and McCLendon (2000) used large random samples of the Dutch (N = 2,099) and French (N = 1,258) speaking population in Belgium. They tested a measurement model for two balanced sets of items measuring two related concepts (the attitude towards immigrants and political distrust) and showed that a model with an additional style factor fitted the data much better than a model without that style factor. The factor loadings for the style factor were all positive and significantly different from zero, but relatively small and could be set equal for all Likert-items in the two cultural groups. The variance of the style factor was much smaller than the variances of the content factors. Hence, they behaved exactly as could be expected of a method (or style) effect. After all, it seems

[8] The thresholds were derived from PRELIS2 computation of polychoric correlations on the joined observations of the countries (Jöreskog, 1993).

[9] During the discussions about this issue at the Large Scale Facilities conference in Cologne, Albert Satorra too defended this position.

plausible that the size of a method effect is not as large as that of a substantive effect. It also seems quite reasonable that an agreeing-response bias would affect all items of a battery of Likert-type items in the same manner. The authors also showed that in the two cultural groups, the style factor correlated nearly perfectly ($r > 0.90$) with a variable that was constructed as the number of times the respondent agreed to the 14 items of a balanced scale (this is called "scoring for acquiescence"). This agreeing-response style can be considered as a method effect because it is caused by the Likert-format of the items (agree-disagree).

Table 3: Summary statistics for the models with one content factor.

Country	Chi-square	df	RMSEA	p-value of close fit	Var (F) (t-ratio)
Austria	244.290	27	0.100	0.000	0.375 (6.816)
(West) Germany	307.109	27	0.101	0.000	0.583 (11.733)
Ireland	215.085	27	0.0905	0.000	0.213 (5.095)
Italy	159.076	27	0.0706	0.000	0.226 (6.021)
The Netherlands	129.885	27	0.0651	0.009	0.227 (6.826)
Norway	152.066	27	0.0591	0.045	0.281 (8.348)
Spain	121.989	27	0.0583	0.086	0.178 (5.986)
Sweden	149.301	27	0.0649	0.006	0.299 (7.690)
United Kingdom	195.652	27	0.0851	0.000	0.475 (9.499)

In this paragraph, it is evaluated whether the inclusion of the method effect into a model with only a substantive factor leads to an improvement of the model fit. For each country, the fit of a model without a style factor is compared to a model with a style factor. Table 3 shows some summary statistics for the models with one substantive factor, but without a style factor. All these models contain the nine items that were described in Table 1. These items are assumed to measure one latent trait, namely: feeling threatened by immigrants on an economic and a cultural level. Theoretically, one could expect two separate dimensions, but previous research with comparable items about economic and cultural threat showed one general dimension (Billiet, 1996; Verberk, 1998: 161). Moreover, one substantive factor was also found in the sample (the Netherlands) that was used for exploration. Except for the first indicator λ_{11}, which was fixed to 1 for scaling reasons, all other factor loadings of the items on the substantive factor are unconstrained (λ_{i1} = free for $i \in \{2,\ldots,9\}$), as was the variance of the content factor.

Obviously, the models in Table 3 don't fit the data. However, whether the models fit the data or not is currently not relevant. For now, we only want to evaluate whether the inclusion of a style factor in these models leads to a significant decrease of the Chi-square statistic for these models. Therefore, we will compare the Chi-square statistic of the models from Table 3 to that of the models that include a method factor next to the substantial factor.

As in our previous research on acquiescence, all loadings λ_{i2} on the method factor are constrained to be equal (except item λ_{82}), because of the assumption that all Likert-type items are equally susceptible to the agreeing response bias.[10] The factor loading on the style factor of the eighth item (Q51) is not constrained (λ_{82} free) because that item is not susceptible to the acquiescent response bias. We expect that the loading of this item on the style factor will not be significantly different from zero. The covariance between the substantive factor and the style factor was set to zero (cov(F,M) = 0) because it is very unlikely that the agreeing-response bias is related to the substantive latent variable.

Table 4: Summary statistics for the models with one substantive factor and a style factor.

Country	Chi-square	df	RMSEA	p-value of close fit	Var(F) (t-ratio)	Var(M) (t-ratio)
Austria	108.689	26	0.0590	0.0893	0.379 (6.978)	0.078 (8.845)
(West) Germany	172.143	26	0.0719	0.00018	0.576 (11.797)	0.056 (9.104)
Ireland	113.331	26	0.0598	0.0684	0.241 (5.522)	0.055 (8.136)
Italy	94.490	26	0.0499	0.484	0.231 (6.169)	0.052 (6.758)
The Netherlands	47.808	26	0.0290	0.998	0.260 (7.189)	0.036 (6.873)
Norway[a]	71.000	25	0.0371	0.983	0.284 (8.452)	0.019 (4.893)
Spain	80.744	26	0.0444	0.788	0.171 (5.879)	0.031 (5.642)
Sweden	124.344	26	0.0604	0.0425	0.307 (7.819)	0.021 (4.510)
United Kingdom	85.668	26	0.0478	0.602	0.479 (9.605)	0.046 (8.235)

[a] Because of identification purposes the error covariance between item 3 and item 6 (cov ($\varepsilon_{6,3}$)) was not constrained at zero

[10] This assumption is not necessary for our arguing. It is possible that some items are less susceptible to acquiescence than other ones depending on the content of the items.

On the basis of a comparison of Tables 3 and 4, we can conclude that the specification of a style factor that is significantly different from zero always results in a very significant drop in the Chi-square statistic. Hence, we may conclude that in all countries, the Likert-items are susceptible to an agreeing-response bias and that models that specify the method factor perform much better than models without a method factor. The method factor behaves in precisely the same manner as was found in the Belgian data. After the response scales were inverted in order to assign the maximum score (5) to the answering category 'completely agree', all the factor loadings on the method factor have small, significant, positive and equal loadings (except λ_{82}) in nearly all countries. The method variance (var(M)) is always substantially lower than the variance of the content factor (var(F)). Finally, the style factor correlates very strongly with a variable 'sum of agreements'. Hence, we may conclude that the method factor is the tendency to agree with Likert-items.

The method that was developed with data of two Belgian regions can be generalised to a substantial number of Western European countries. The next section, which evaluates the construct equivalence of the substantive factor, will also investigate whether the method effect differs from one country to another. If this is the case, we would have a method bias in the cross-cultural sense.

4 Evaluating the construct equivalence of the content factor

4.1 Model specifications for multigroup analysis

Construct equivalence requires that the construct is operationalized in a sufficiently similar manner over the countries. Rensvold and Cheung (1998) define construct equivalence operationally as factorial invariance. This means that in the multigroup analyses equality constraints will be placed on the factor loadings of the substantive factor, in the sense that they are set equal across the countries that are included in the analysis. This is indicated by equation 1.

$$\lambda_{i1}^{1} = \ldots = \lambda_{i1}^{k} \text{ (for } i \in \{1,\ldots,9\} \text{ and } k = \text{number of countries)} \qquad (1)$$

Hence, a construct is not equivalent if the modification indices indicate that the equality constraints on the factor loadings of the substantive construct have to be relaxed. There are several options when noninvariant items are detected (the construct has a substantially different meaning in some groups) (Rensvold and Cheung, 1998). The first is to exclude at least one of the countries from the scale. For this country, then, the construct is not equivalent to that of the other countries. Second, one can remove the items for which the equality constraint has to be relaxed from the multigroup analysis. The revised scale, then, might be equivalent across all countries. One has to take care, however, not to remove too many items from the scale because this would lead to a poorly represented construct (i.e. construct underrepresentation). Third, one can accept the idea of partial factorial invariance (Byrne et al., 1989). Invariant indicators may be retained if the researcher can argue that they have an insignificant effect on the outcome of the analysis. Finally, the failure to assess complete invariance can lead to the conclusion that the construct has a substantially different meaning in some groups (Rensvold and Cheung, 1998).

As in the previous paragraph, we assume that each of the positively and negatively worded indicators are equally susceptible to the acquiescent response bias and that the eighth item (Q51) is not affected by this type of bias:

$$\lambda_{12}{}^{1} = \ldots = \lambda_{i2}{}^{k} \quad (\text{for } i \in \{2,\ldots,7,9\}) \qquad (2)$$

We also assume that there is no difference between the countries with respect to the agreeing-response bias (cfr. equation [3]); in other words: the tendency to agree with items regardless of their content is expected to be independent of the country the respondents belong to. After all, we only included West-European countries in our analysis.

$$\lambda_{i2}{}^{1} = \ldots = \lambda_{i2}{}^{k} \quad (\text{for } i \in \{1,\ldots,9\}) \qquad (3)$$

Equation [3] also implies, however, that we hypothesise that the acquiescent response bias is not a source for a cross-cultural method bias. It will, however, be tested whether the hypothesis of no cross-cultural method bias due to differences in susceptibility to the agreeing-response bias holds. As before, no restrictions were placed on the variances of the latent constructs; the covariance between the substantive construct and the style factor is fixed at zero.

4.2 Determining where to start

When looking for groups of countries with equivalent constructs within a population of 9 countries, it is important to know where to start. In other words, it is important to know which countries should be combined first and where to go from there. After all, if one starts to combine countries haphazardly and tries to form equivalent groups by trial and error, chances are that one will never find the most optimal combination of countries. Hence, a more formal way of deciding which countries to combine, seems necessary.

One way to decide in which way the countries can be grouped, is by performing a (hierarchical) cluster analysis on the factor loadings for the content factor. Hence, in this analysis the factor loadings of the indicator variables on the latent construct serve as the classification variables and the countries are the observations that have to be combined into groups. However, there are still several possible ways to perform a hierarchical cluster analysis. Here, we will use four different clustering methods: single linkage clustering, complete linkage clustering, average linkage clustering, and the method of Ward.[11] We will use the Euclidean distance as a measure for the similarity between the observations. The results of the cluster analyses are shown graphically in Figure 1.

[11] Single linkage clustering might be less appropriate for the current analysis because of its tendency towards ‘chaining’.

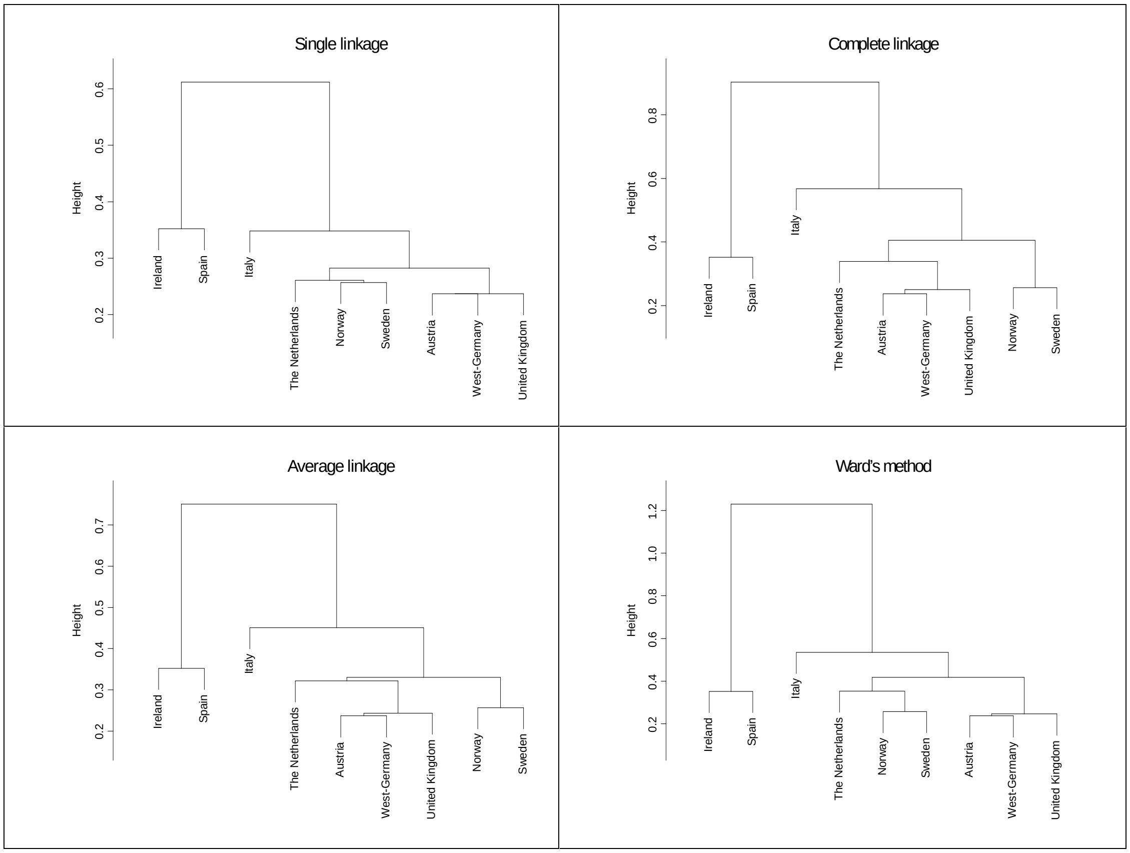

Figure 1: Clustering trees for the four hierarchical clustering methods.

From this figure, it can be seen that the results of all four clustering methods are quite similar. Every method first combines Austria, West Germany and the United Kingdom into one group. Afterwards, all methods merge Norway and Sweden into a new cluster. From this point on the results of the four methods are different. The single linkage method and Ward's method put the Netherlands into a new cluster together with Norway and Sweden, whereas the other two methods classify the Netherlands into a cluster with the United Kingdom, West Germany and Austria. Another recurrent finding in all of the analyses is that Spain and Ireland are quite similar to each other, but very different from the other countries.

4.3 Results of the multigroup analysis

According to the results of the cluster analysis, the 9 West-European countries can be divided into two different groups. One cluster of countries consists of The Netherlands, Norway, Sweden, Austria, (West-) Germany, and the United Kingdom. The other cluster contains Ireland and Spain. Italy is neither closely related to the first group, nor to the second group. We decided to include this country in the cluster with Ireland and Spain because of theoretical reasons. In a further step of this research project, we are interested in the relationship between ethnocentrism and religion. From that point of view, it is useful to build a group of strongly catholic denominated countries.[12]

Cluster 1: The Netherlands, Norway, Sweden, Austria, (West) Germany, and the United Kingdom.

The positively worded item about the economy (« Immigrants are generally good for [country's] economy » (Item 5) showed very low factor loadings for most countries (especially for Sweden). Therefore that item was deleted and further analyses were only conducted on the eight remaining items.

The test started with a completely constrained model. The variances of the two factors (content factor and style factor) are constrained to be equal over all groups. As was specified in the equations [1], [2] and [3], the factor loadings on both the content factor and the style factor were set equal between the countries and the factor loadings of the method factor are all constrained to be equal within the countries. The structure of the error covariances was allowed to differ between the groups. This means that some differences in error covariances are considered not to affect the meaning of the measured construct in a substantial manner. In sum, only seven error covariances were not fixed to be zero (none in The Netherlands and the UK, 1 in Norway and Austria, 2 in Germany and 3 in Sweden). Some goodness of fit statistics of the constrained model are reported in Table 5 (see Model 2). According to our criteria, this model is acceptable (RMSEA < 0.05; the p-value of close fit equals 1; NFI is close to 1). Remember that we have placed very strong constraints on the model parameters, namely complete factorial invariance between the groups for the factor loadings of the content factor and complete factorial invariance within and between the groups for the factor loadings of the method factor. It is certainly possible to improve the model somewhat by relaxing some other error covariances and some λ_{i1}'s in some countries.

In order to have a better idea about the amount of possible improvements, the completely constrained model is compared with the semi-constrained model (Model 3). This is a model in which the method factor is factorially invariant within and between the groups, but in which the loadings of the content factor are variant (completely unconstrained). The difference in Chi-square between the constrained and the unconstrained model divided by the difference in degrees of freedom gives an idea of the average improvement of the fit that can be obtained for each free factor loading. This is equal to 5.72 per degree of freedom, which means that a better model can be obtained by relaxing some of the λ_{i1}'s (see Table 5).

The goodness of fit statistics for the constrained model without a method factor (Model 1) is also included in Table 5. This provides an idea about the improvement that we have realized by specifying an additional method factor. This improvement is considerable (a drop of 26.9 Chi-square units per degree of freedom).

[12] The project is sponsored by the National Research Fund, Flanders. Grant nr. G.0125.98.

Table 5: Some goodness of fit statistics for the constrained model without a method factor (Model 1), the constrained model with content and method factor (Model 2), and the semi constrained model (Model 3): six countries.

Model	Chi-Square	Df	RMSEA	p-value (close fit)	NFI
(1) Content : constrained	740.592	148	0.0626	1.0	
(2) Content and Method : constrained	417.509	136	0.0434	1.0	0.978
Drop in Chi-square = -323.08 for 12 df (drop of 26.9 units per df)					
(3) Content and Method : semi-constrained	274.484	101	0.035	1.0	0.983
Drop in Chi-square = -143.03 for 35 df (drop of 4.09 units per df)					

The λ_{ij} parameters estimated under Model 2 are reported in Table 6. Only one set of factor loadings on the content and the method factor is reported since they are invariant over the six countries. The correlations of all items with the content factor are strong. The loading of almost 0.20 on the method factor demonstrates that there exists a significant method effect in the six countries. The small significant loading of the neutral item Q51 on the style factor for Germany is probably due to chance.

Table 6: Parameters estimated under Model 2: factorial invariant for content and method factors in The Netherlands, Norway, Sweden, United Kingdom, Austria, and (West-) Germany. Common metric standardised solution (t-values for free parameters within brackets).

Items	Content Factor (Outgroup)	Method factor (Acquiescence)
Q42	0.588 (fixed)	0.198 (fixed)
Q44	0.648 (34.866)	0.198
Q45	-0.566 (-30.194)	0.198
Q47	0.776 (37.938)	0.198
Q49	0.729 (37.449)	0.198
Q50	-0.648 (-34.009)	0.198
Q51	0.763 (38.200)	ns[1]
Q52	-0.671 (-33.492)	0.198

[1] Not significantly different from zero in all countries, except for Germany (λ_{82}^{6} = 0.099 ; t-value 2.172).

Since we have reversed the codes (‘completely agree’ = 1 ; ‘completely disagree’ = 5), the positively worded items now have negative signs on the content factor and the negatively worded items have positive signs. This means that the higher the scores on the latent variable, the more negative the attitude towards immigrants.

The variances of the latent variables are nearly the same as reported in Table 4. For that reason we do not repeat them here. Remember that the variances of the method factor are significantly different from zero (t > 3.2 for all countries).

By controlling for acquiescence, it is possible to obtain a set of valid indicators for the content factor that are invariant over the six countries. The variance of the construct that measures feelings of threat towards immigrants is no longer biased by the tendency to agree with the statements. The construct equivalence achieved here allows researchers to compare the level of ethnocentrism in these countries by comparing the means of the latent variable. In structural models, the correlations with other variables are also ‘cleaned’ for the agreeing response style.

Researchers who want the Chi-square value to be no larger than 3 times the degrees of freedom (Bollen, 1989: 278; Carmines and McIver, 1981)may argue that it would be better to select a model that meets this criterion. Is this possible without falling back on the semi-constrained model where all the loadings on the content factor are unconstrained? When some loadings on the content factor (λ_{i2}’s) are freely estimated in one or two of the countries, one can obtain a variant of Model 2 that satisfies this criterion (Chi-square = 305.91; df = 127). This is a substantial improvement of the fit.

The parameters of this new Model 2b are shown in Table 7. The loadings of all the Likert-type items on the method factor are still invariant in all countries. The loading of indicator Q51 on the style factor in the German sample is no longer significantly different from zero at the 0.05 level, as is the case in the other coutnries. Three of the eight indicators continue to have invariant loadings on the content factor in all six countries. Two indicators show a variant loading in one group, whereas the remaining two indicators have variant factor loadings in two groups. Here we have to ask whether the construct still has the same meaning in all six groups? Answering this question requires careful investigation by cultural specialists who know the different languages. From a pure technical point of view, one can see that there is a shift in the importance of the items. In Model 2, the underlying construct was most strongly related to the items about crime (Q47), the number of immigrants (Q51), and the loss of jobs (Q49). This is still the case in most groups, but the item about customs and traditions (Q44) is now somewhat more important in the Netherlands. In Norway, the items concerning new ideas and cultures (Q50) and number of immigrants are now more strongly related to the construct. The issue of buying land (Q42) became a more important indicator in the United Kingdom, while in West Germany the item about customs and traditions (Q44) is the most important item. All in all, we have the impression that the construct continues to have approximately the same meaning in the different groups. Next, we will evaluate the similarity of the construct for the three remaining countries.

Table 7: Parameters estimated under Model 2b: partial factorial invariant for content factor in The Netherlands, Norway, Sweden, United Kingdom, Austria, and (West-) Germany. Common metric standardised solution (t-values for free parameters within brackets).

Items	Invariant loadings in at least four countries		Variant loadings on content factor					
	Method factor	Content Factor	The Nethe rlands	Nor-way	Swe-den	United Kingdom	Austria	(West) Germany
Q42	0.199	0.540				0.701		0.686
Q44	0.199	0.603	0.730			.		0.814
Q45	0.199	-0.565						
Q47	0.199	0.780						
Q49	0.199	0.758			0.628			
Q50	0.199	-0.626		-0.729				
Q51	n.s.	0.761		0.848			0.669	
Q52	0.199	-0.669						
Chi-square = 316.859; df = 128; RMSEA = 0.0368; p-value of close fit = 1.0; NFI = 0.980								

Cluster 2: Ireland, Spain, and Italy.

Achieving construct equivalence for the three countries belonging to the second cluster was far more difficult. This could be expected because of the extension of the cluster formed by Spain and Ireland with Italy. It was impossible to retain more than six items. Three items had very low factor loadings : the two items about customs and traditions (Q44 and Q45) and the non-Likert item (Q51) concerning the increase or decrease of immigrants. The positively worded item about the economy (Q48), which was not retained in the previous cluster is now added to the list of indicators. Because of this, the construct covers another content domain than the construct in the previous cluster (with six countries). The constrained model (Model 2) with two factors (content and method), again yields a considerably better fit than the model with only a content factor (Model 1). This time, the semi-constrained model does not lead to a substantial improvement of the fit. (This is, however, also due to the fact that three items were already removed from the scale.). As a consequence, we can retain the model with complete factorial invariance, but because of the dropping 3 items, the content domain is now much narrower than in the other cluster. That is the price that had to be paid for obtaining an equivalent construct for the three countries in the cluster.

Because the number of indicators changed considerably, the variances of both the content and the style factor are somewhat different from those reported in Table 4. Most important, in Italy the variance of the method factor is no longer significant (var(M) = 0.005 ; t = 0.438)). This means that there is no detectable agreement bias in Italy. The parameters of the constrained two factor model (Model 2) are reported in Table 9.

Table 8: Some goodness of fit statistics for the constrained model without a method factor (Model 1), the constrained model with content and method factor (Model 2), and the semi constrained model (Model 3): three countries (Spain, Italy, and Ireland).

Model	Chi-Square	Df	RMSEA	p-value (close fit)	NFI
(1) Content : constrained	95.065	33	0.043	1.0	0.963
(2) Content and Method : constrained	70.058	30	0.0358	1.0	0.973
Drop in Chi-square = -25.01 for 3 df (drop of 8.34 units per df)					
(3) Content and Method: semi-constrained	43.378	20	0.033	1.0	0.983
Drop in Chi-square = -26.68 for 10 df (drop of 2.67 units per df)					

Table 9: Parameters estimated under Model 2: factorial invariant for content and method factors in Spain, Italy, and Ireland. Common metric standardised solution (t-values for free parameters within brackets).

Items	Content Factor (Outgroup)	Method factor (Acquiescence)
Q42	0.473 (fixed)	0.134
Q47	0.638 (16.201)	0.134
Q48	-0.497 (-14.759)	0.134
Q49	0.703 (16.407)	0.134
Q50	-0.488 (-14.809)	0.134
Q52	-0.446 (-13.771)	0.134

Considering the strength of the correlations with the underlying content factor, we must conclude that the measurement quality of the construct is much lower in cluster 2 (Italy, Spain, Ireland) than in the first cluster (The Netherlands, Norway, Sweden, United Kingdom, Austria, and (West-) Germany).

5 Conclusion

The scope of this article was to evaluate whether a battery of nine items which were supposed to measure the same latent trait in nine West European countries, actually did measure the same construct. Structural equation modelling was used to tackle this problem. This analysis technique was preferred over other techniques, because it allows us to control for a specific type of method effect, called acquiescence. It was expected that this type of method bias affected the answers to the eight Likert-type items of the scale. Therefore, we first evaluated whether including a style factor in the models to account for this agreeing-response bias lead to a better fit to the data. The results clearly show that in each of the countries the scale that measures the respondent's feelings of threat towards immigrants is susceptible to an agreeing-response bias. Hence, the data are influenced by an intracultural method effect.

Next, it was investigated whether construct equivalence could be achieved for all nine countries and for the entire scale of nine items. Construct equivalence was operationally defined as factorial invariance. A cluster analysis on the factor loadings of the countries on the content factor showed that complete factorial invariance could not be achieved. It suggested that two distinct clusters of countries can be retained. The first cluster contains Austria, The Netherlands, Norway, West Germany, Sweden and the United Kingdom. The second cluster consists of Spain and Ireland.

We described four ways of dealing with factorial non-invariance. The first was to remove at least one of the countries from the analysis. This is in line with our way of working. After all, on the basis of the cluster analysis we decided to divide the entire group of nine countries into two subgroups (Italy was included into the cluster with Spain and Ireland for theoretical reasons). A second way of dealing with deviations from factorial invariance is to remove certain items from the scale. We also applied this procedure to our data. In the first cluster only eight out of nine items were used. In the second cluster, three items had to be removed from the scale to obtain factorial invariance. Thirdly, construct equivalence can also be obtained if some of the equality constraints on the factor loadings can be relaxed without substantially changing the outcome of the analysis. This is not necessary for the second cluster (with Spain, Italy and Ireland). For the first cluster, however, this leads to a better fitting model in which the overall meaning of the construct does not seem to differ between countries, although in some countries there is a stronger emphasis on some items than in other countries.

The analyses also show that the models that include a style factor always fit the data better than the models without a style factor. Thus, the agreeing-response bias that was detected by Billiet and McClendon (1998,1999) in a sample of Belgian respondents can be generalised to a substantial number of West European countries.

With respect to this acquiescent response bias, we also hypothesised that it would not be a source of cross-cultural method bias. In other words: we assumed that the people of the nine West European countries involved in the analysis, were equally prone to the response bias. This hypothesis seems to receive support from the data. Most importantly, the equality constraints on the factor loadings of the style factor (within as well as across the countries) never had to be relaxed. In the second cluster of countries, however, the variance of the style factor was not significantly different from zero for the Italian sample. Hence, the agreeing-response style could not be shown to influence the Italian respondents. Secondly, the models did not test whether the factor loadings on the style factor are the same over the two clusters of countries, although this seems reasonable in the light of the small differences between the two clusters (0.198 versus 0.134). Thus, it seems reasonable to conclude that the agreeing-response bias does not lead to a cross-cultural method bias.

However, there are other potential sources of a cross-cultural method bias. First of all, the surveys were conducted by different research facilities. It is not unlikely that this leads to differences in the administration of the questionnaire, etc. Secondly, there are obvious differences in the sampling schemes of the countries. The Austrian sample, for instance, is representative of the Austrian population of 14 years and older. The West German and the British sample, however, only include respondents of 18 years and older, whereas the Norwegian sample is representative of the population between 16 an 79 years old. These differences in the sampling scheme do not allow a direct comparison of the scores of respondents from different countries on the scale. However, they don’t affect the construct equivalence of the scale (van de Vijver and Leung, 1997).

References

[1] Berry, J.W., Poortinga, Y.H., Segall, M.H., and Dasen, P.R. (1993): *Cross-Cultural Psychology. Research and Applications*. Cambridge: Cambridge University Press.

[2] Billiet, J. (1996): *Theoretical Dimensions and Measurement Models of Attitudes towards Ethnic Minorities.* Leuven: ISPO-bulletin 1996/23.

[3] Billiet, J.B. and McClendon, M.J. (1998): On the identification of acquiescence in balanced sets of items using structural models. In A. Ferligoj (Ed.): *Advances in Methodology, Data Analysis and Statistics.* Ljubljana: FDV.

[4] Billiet, J.B. and McClendon, M.J. (2000): Modeling aqcuiescence in measurement models for two balanced sets of items. *Structural Equation Modeling,* **7**, 608-628.

[5] Bollen, K.A. (1989): *Structural Equations with Latent Variables*. New York: Wiley.

[6] Bollen, K.A. and Long, J.S. (1992): Tests for structural equation models. *Sociological Methods and Research*, **21**, 123-131.

[7] Byrne, B.M., Shavelson, R.J., and Muthén, B. (1989): Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, **105**, 456-466.

[8] Carmines, E.G. and McIver, J.P. (1981): Analyzing model with unobserved variables: Analysis of covariance structures. In G. W. Bohrnstedt and E. F. Borgatta (Eds.): *Social Measurement: Current Issues*. Beverly Hills: Sage.

[9] Coenders, G. (1996): *Structural Equation Modeling of Ordinal Data*. Barcelona: ESADE (Unpublished PhD dissertation).

[10] Coenders, H. and Saris, W. (1995): Categorization and measurement quality. The choice between Pearson and Polychoric correlations. In W. Saris and A. Munnich (Eds.): *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments*. Budapest: Eötvos University Press.

[11] Homer, P. and Creech, R.M. (1988): Using LISREL Models with Crude Rank Category Measures. *Quality and Quantity*, **322**, 191-202.

[12] Hui, C.H. and Triandis, H.C. (1985): Measurement in cross-cultural psychology. A review and comparison of strategies. *Journal of Cross-Cultural Psychology*, **16**, 131-152.

[13] Johnson, D.R. and Creez, J.C. (1983): Ordinal measures in multiple indicator models. A simulation study of categorization error. *American Sociological Review*, **48,** 398-407.

[14] Johnson, T.P. (1998): Approaches to equivalence in cross-cultural and cross-national survey-research. *ZUMA Nachrichten Spezial, Cross-Cultural Eurvey Equivalence*, **3**, 1-40.

[15] Jöreskog, K.G. (1990): New developments in LISREL. Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, **24**, 387-404.

[16] Jöreskog, K.G. (1993): *New Features in PRELIS2*. Scientific Software Intenational.

[17] Likert, R. (1932): A Technique for the Measurement of Atittudes. *Archives of Psychology*, **40**.

[18] O’Brien, R.M. (1985): The relationship between ordinal measures and their underlying values: Why all the disagreement? *Quality and Quantity*, **19**, 265-277.

[19] Rensvold, R.B. and Cheung, G.W. (1998): Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement*, **58**, 1017-1034.

[20] Van De Vijver, F.J.R. (1998): Towards a theory of bias and equivalence. *ZUMA Nachrichten Spezial, Cross-Cultural Survey Equivalence*, **3**, 41-65.

[21] Van De Vijver, F. and Leung, K. (1997): *Methods and Data Analysis for Cross-Cultural Research*. Thousand Oaks: Sage.

[22] Verberk, G. (1998): *Attitudes towards Ethnic Minorities. Conceptualisations, Measurements, and Models*. Nijmegen: Published doctoral dissertation.

[23] Welkenhuysen-Gybels, J. (1998): Cross-cultureel onderzoek: methodologische problemen bij de toepassing van een verkorte F-schaal in Franstalig en Nederlandstalig België. *Tijdschrift voor Sociologie*, **19**, 449-472.