A Comparison of Methods for the Evaluation of Construct Equivalence in a Multigroup Setting

Jerry G.J. Welkenhuysen-Gybels, FWO-Flanders/ Catholic University of Leuven, Belgium
Fons J.R. van de Vijver, Tilburg University, The Netherlands

Keywords: Cross-cultural research, Construct equivalence, Factorial invariance

INTRODUCTION

Cross-cultural survey research, whether it involves a comparison of cultures, nations or language groups, usually has to deal with more methodological issues than an intracultural survey (Van de Vijver 1998; Van de Vijver & Leung, 1997a,b). Perhaps the most prominent is the problem of making valid comparisons across cultures. The comparability of these scores depends on their level of equivalence.

In the literature several types of equivalence have been defined (Johnson, 1998). Her we focus on construct equivalence (van de Vijver 1998; van de Vijver and Leung, 1997a, b). Construct equivalence implies that respondents from different cultural groups attach the same meaning to the construct as a whole. Several methods have been proposed for evaluating construct equivalence, which typically consist of a pairwise comparison of factors or dimensions across cultural groups (e.g., van de Vijver and Leung 1997a, b).

Applying these techniques in studies involving a large number of groups (e.g., Inglehart 1993, 1997; Schwartz 1992) gives rise to two related problems that are addressed in the present paper. The first one is primarily technical and involves the rapid growth of pairwise comparisons of factors across groups when the number of groups is large. The second is more conceptual; in a study involving many groups it is quite likely that not all cultural groups will have equivalent constructs. The problem to deal with is to identify homogeneous partitions of the cultural groups that show construct equivalence within the cluster. The present paper proposes three different search strategies for dealing with the two problems mentioned. These are discussed in the next section. Some resampling methods to determine critical values for two of these procedures are also introduced. In the third section, these strategies are applied to the data of 1990-1991 World Values Survey (Inglehart 1993, 1997). The paper ends with a comparison and a discussion of the results of the three approaches and their resampling methods.

METHOD

TECHNIQUES FOR EVALUATING CONSTRUCT EQUIVALENCE

In the current paper, we only discuss the application of exploratory factor analysis to the assessment of construct equivalence, because it is a relatively simple technique that is widely available in software packages such as SAS and SPSS. In exploratory factor analysis construct equivalence is defined operationally as factorial invariance (Meredith 1993; Rensvold and Cheung 1998; ten Berge, 1986). This definition implies that a construct is equivalent across cultural groups if the factor loadings of the items on the latent factor are invariant across cultural groups. The agreement between the factor loadings of items from two different groups can be expressed via Tucker's phi (Tucker, 1951). The index measures the identity of two factors, up to a positive, multiplying constant. The latter allows for differences in factorial eigenvalues across cultural groups. Unfortunately, the index has an unknown sampling distribution, which makes it impossible to construct confidence intervals. Some rules of thumb have been proposed: values higher than 0.95 are taken to indicate factorial invariance, whereas values lower than 0.90 (Van de Vijver and Poortinga 1994) or .85 (Ten Berge 1986) point to nonnegligible incongruities. As an alternative, Chan, Ho, Leung, Chan and Yung (1999) have proposed a bootstrap procedure to determine a critical value for these congruence indices when two groups are to be compared. The next section discusses how exploratory factor analysis can efficiently be applied in a multiple group context and develops multiple group resampling methods to determine critical values for the agreement indices.

To avoid aggregation errors in hierarchically structured data (cross-cultural data by definition have such a structure) a within-subgroups standardisation will be performed on the data (Muthén 1991, 1994).

METHODS FOR EVALUATING CONSTRUCT EQUIVALENCE IN A MULTIGROUP SETTING

Top-down approach. A first approach in the quest for equivalent partitions of the groups under study, is a top-down approach. At the start of this approach, the studied groups are all combined into one set. This set is referred to as the pooled data set. The pooled data set treats the groups as if they were all equivalent and originate from the same population (after correcting for differences in mean scores and standard deviations across cultures). If all groups originate from the same population, the agreement between the individual groups and the pooled data set should be very high. A low agreement between the pooled data set and a particular group, on the other hand, indicates that this group does not belong to the same parent population as the other groups and should therefore be removed from the pooled data set. The group with the lowest agreement to the pooled data set is removed first. A new, slightly smaller pooled data set is constructed from the remaining groups and the similarity of the constituting groups to this data set can be calculated again. As before, the group that is least similar to the pooled data set has to be removed. This process continues until the similarities between the pooled data and the individual groups that constitute the pooled data set are all above a certain critical value (which will be determined via a resampling procedure, cfr. infra). The groups that have been removed from the original pooled data are then pooled so as to start a new series of evaluations of factorial agreement. Such a repeated

application can be expected to lead to partitions of equivalent groups.

Bottom-up approach. The bottom-up approach addresses the problem from the opposite direction. It starts with a matrix of pairwise agreement indices for all groups in the study. From this matrix, the two groups that are the most similar are combined to constitute the pooled data set. Next, the agreement indices between the remaining groups and the newly constructed pooled data set are calculated and the group with the highest agreement to the pooled data is added to the latter. This iterative process continues until none of the remaining groups has an agreement index that is above a certain critical value. The remaining groups are then again scrutinized for the two most similar countries and the bottom-up process starts again for these remaining groups.

Critical values for the bottom-up and the top-down approach. The previous section mentioned different ways to determine a critical value for the Tucker's Phi. Here, multigroup resampling methods are proposed to determine the critical values for Tucker's Phi, which are generalisations of the bootstrap procedure for the two-group case, developed by Chan et al. (1999).

We first consider the top-down approach. Suppose that the pooled data set consists of m groups with sample sizes $n_1,\ldots, n_m$, respectively. The resampling procedure is as follows:

a) Select m random samples of sizes $n_1,\ldots, n_m$, respectively from the pooled data set.
b) Compute the agreement (via Tucker's Phi) of the factors obtained in the m random samples and the pooled data set.
c) Repeat the two previous steps as many times as required.

The obtained sampling distribution is the distribution of the agreement index if the groups under study would be random samples from the same population, namely the pooled data set. Groups with an agreement index that is smaller than the critical quantile of this sampling distribution, can be conceived as not being drawn from the general population and should be removed from the pooled data set.

In the bottom-up approach the groups for which the sampling distribution has to be obtained, are by definition not part of the pooled data set and a different resampling method has to be used than in the top-down approach. Suppose that the pooled data set has a sample size of N, and that there are k groups, with respective sample sizes $n_1,\ldots, n_k$, that are eligible to be added to the pooled data set. The procedure for the bottom-up approach becomes:

a) Combine the pooled data set and the data set with groups that are eligible to be added to the pooled data set. We will refer to this data set as the joint data set.
b) Select $k + 1$ random samples of size N, $n_1,\ldots, n_k$, respectively from the joint data set.
c) Compute the agreement (via Tucker's Phi) between the random sample of size N (random pooled data set) and the k random samples of size $n_1,\ldots, n_k$, respectively.
d) Repeat steps b-c as many times as required.

A few remarks have to be made with regard to the proposed resampling procedures. First of all, the resampling procedure for the top-down approach should not be applied in every step. If it were applied in every step, this could easily lead to exactly the opposite of what one wants to achieve (i.e. partitions of equivalent groups). Removing the least similar group from the pooled data set makes the latter data set more homogeneous. As a consequence, resampling from the new pooled data set yields a higher critical value for the agreement index than in the previous step. If resampling were performed in each step of the top-down procedure, we might end up with a data set that contains only a single or at least very few groups.

To avoid this problem, we compute a critical value for the agreement index via the proposed resampling method at the start of the procedure and keep using this critical value until all groups in the pooled data set have an agreement index that is higher than the critical value and no more groups have to be removed from the pooled data set. At the start of the next phase a new critical value for the remaining groups has to be determined. After all, the pooled data set that is used in this next phase is very different from the pooled data set from the previous phase. In the bottom-up approach this problem of unwanted homogenization does not occur and the resampling procedure could in principle be used in each step.

A second remark pertains to the selection of the two most similar groups from a similarity matrix at the beginning of each phase in the bottom-up approach. The question is whether the two most similar groups are similar enough to be combined in the first place and how to determine a critical value for this evaluation. In our view, there are three ways to tackle this problem. First, one can rely on the rules of thumb that were mentioned in the previous section. A second approach would be the usage of the pairwise resampling method developed by Chan et al. (1999), to determine the critical value. This approach is not very attractive either as the number of pairwise comparisons increases rapidly as the number of groups to be compared becomes larger; it is indeed the purpose of the bottom-up and the top-down procedures to avoid this problem. A third possibility, is to use the top-down resampling procedure on the joint data set. This provides a sampling distribution for the agreement between random samples from the joint data set and the joint data set itself. Since the joint data set can be interpreted as an average of the populations in the random samples, top-down resampling seems to constitute a viable alternative to an actual pairwise resampling procedure in the multiple group case.

Heuristic approach. In this last approach, a matrix of pairwise agreement indices between the cultural groups is used as input for some dimension-reduction technique, such as cluster analysis.

ILLUSTRATION

The three procedures from the previous section are illustrated on the basis of the 1990-1991 World Values Survey (Inglehart 1993, 1997). The study involved a total of 47,871 respondents from the following 39 "regions".

Attitudes toward postmaterialism were examined. The inventory comprised of 12 items (reproduced in Table 1) that were presented as three quadruplets. For each quadruplet a card was shown to the respondent with four values printed on it (e.g., the first four items of Table 1). Respondents were asked to rank these items according to their importance to them. The first rated option got a score of 3, the second of 2, and the remaining options received a score of 1.

TABLE 1: Items of the Postmaterialism Scale (Ingelhart, 1993)

Nr	Item	Dimension
1	Making sure this country has strong defence forces	Materialism
2	Seeing that people have more to say about how things are done at their jobs and in their communities	Postmaterialism
3	Trying to make our cities and countryside more beautiful	Postmaterialism
4	Maintaining order in the nation	Materialism
5	Giving people more to say in important government decisions	Postmaterialism
6	Protecting freedom of speech	Postmaterialism
7	A stable economy	Materialism
8	Progress toward a less impersonal and more humane society	Postmaterialism
9	Progress toward a society in which ideas count more than money	Postmaterialism

Note. The materialist items "Maintaining a high level of economic growth," "Fighting rising prices," and "The fight against crime" were not analysed (see text).

The scoring of the inventory introduces linear dependencies among the data. Each set of values presented on a single card gets a sum score of 7 (being the sum of one score of 3, one score of 2, and two scores of 1). So, the whole instrument has three linear dependencies. Factor analysis may not seem the most appropriate statistical technique here, because of the existence of these dependencies and the, by definition, negative intercorrelations of ipsative scores which may influence the results of a factor analysis. Alternative techniques could be used that are not susceptible to these problems such as multidimensional scaling or unfolding techniques. However, a study by Van Deth (quoted in Inglehart 1997, p. 123) has shown that for the Inglehart data these techniques show results similar to factor analysis. Therefore, it was decided to apply the latter. Linear dependencies were reduced by omitting one materialist item of each set of four that were jointly presented (see Table 1 for an overview of the selected items). Materialist items were omitted because Inglehart's thesis primarily involves postmaterialism.

RESULTS

A SAS® macro 'AGREETUCK' [1] was developed to perform the top-down and the bottom-up procedure a well as the suggested resampling procedures on a given data set. The macro uses Tucker's phi (1951) as a measure for the agreement between the factor loadings. As explained previously, the resampling method for determining a critical value for the Tucker's phi was only applied at the start of a new phase in the aggregation (bottom-up) or the disaggregation (top-down) process. The critical value obtained from this resampling procedure was then used throughout the entire phase.

In the bottom-up as well as in the top-down procedure, the number of iterations for the resampling procedure was chosen in such a way that about 2000 estimates for the Tucker's phi were obtained. An alpha level of 0.05 was used to determine a critical value from the obtained sampling distribution.

Results of the top-down approach. The results of the top-down procedure are shown in Table 2. This table shows that, besides Japan, the first set of countries contains most of the Western European and both Northern American countries. The second set of equivalent regions seems to be a heterogeneous mixture, containing some Western European, Latin American and African regions. A third set contains mostly Eastern European countries and Brazil. In the fourth phase of the procedure, eventually only one country (South Korea) remained in the pooled data set. Apparently South Korea is dissimilar from all the other remaining regions. At first, this might seem a counter-intuitive result. One might expect exactly the opposite: if South Korea is so different from the other remaining countries, then why was it not removed first? There are at least two reasons for this result. First of all, the fact that South Korea is dissimilar from the other regions, does not imply that the other regions are similar to each other. Table 2 shows that in Phase 5 through 8, the remaining regions are split up into 4 sets. This explains why South Korea was not the first region to be removed from the pooled data set: the other remaining regions do not form a monolithic bloc. This, however, does not explain why South Korea is the only region left in the pooled data set. The latter is due to the fact that eliminating regions from the pooled data set, inevitably changes the pooled data set, especially when only few and heterogeneous regions are involved. Thus, the instability of the pooled data set when removing regions from a small heterogeneous set, can trigger such results as in the fourth phase of the top-down procedure. Phase 5 through 7 yield small and heterogeneous sets of regions. Poland is a clear outlier. In every phase, it was the first region to be removed from the pooled data set. Its Tucker's phi for the agreement with the pooled data set from the first phase (which contained all regions) was only 0.58, whereas the Tucker's phis for the other regions were around 0.90 or higher (van de Vijver and Poortinga, in review). Table 2 also illustrates another feature of the top-down approach that may turn out to be a fairly typical feature: larger sets of countries with a fairly homogeneous set of factor loadings tend to be extracted first as they have a relatively large impact on the pooled set because of their number, while single countries ('outliers') are retrieved in a later stage.

TABLE 2: Results of the top-down procedure

Phase	Members (in alphabetical order)	Critical value
1	Austria, Belgium, Canada, Denmark, France, Italy, Japan, The Netherlands, Northern Ireland, Norway, Spain, Sweden, U.S.A., West Germany	0.9817
2	Chile, Iceland, Ireland, Mexico, Nigeria, Portugal, Russia, South Africa, United Kingdom	0.9736
3	Belarus, Brazil, Bulgaria, Czechoslovakia, Estonia, Hungary, Latvia, Lithuania, Moscow	0.9592
4	South Korea	0.9704
5	East Germany, India	0.9589
6	China,Turkey	0.9527
7	Finland	0.8389
8	Poland	--

To determine what distinguishes the sets that were found for the top-down approach, the pooled factor loadings for the first three clusters and Poland are displayed in Table 3.

TABLE 3: Pooled factor loadings for clusters 1-3 of the top-down approach and Poland

Item	Cluster 1	Cluster 2	Cluster 3	Poland
1	-0.31	-0.34	-0.36	-0.44
2	0.55	0.65	0.64	0.72
3	0.13	-0.13	-0.24	-0.53
4	-0.67	-0.70	-0.80	-0.71
5	0.53	0.65	0.77	0.78
6	0.34	0.27	0.27	0.04
7	-0.69	-0.58	-0.38	0.28
8	0.59	0.51	0.43	-0.22
9	0.48	0.42	0.29	0.27

Note. *In order to make loadings comparable across clusters, the loadings of the last three columns have been multiplied by a constant so as to equate the eigenvalues of the factor across the clusters.*

The differences between the first two clusters are relatively small, but meaningful. Whereas in the first cluster (with affluent countries such as Canada, France, Italy, the U.S.A., and West Germany) there is more emphasis on progress towards a humane society in which ideas count more than money, in the second cluster (with countries such as Chile, Mexico, Russia, South Africa, and the United Kingdom) the need for more say in decisions on the job and by the government is slightly more emphasized. In the third cluster (mainly consisting of Eastern European countries) making the countryside more beautiful is a materialist item, there are relatively high loadings for getting more say in decisions and relatively low loadings for progress towards a more humane and a society in which ideas count more than money. On the other hand, materialism is mainly characterized by maintaining order in this cluster. The third cluster seems to have fewer strong indicators for both materialism and postmaterialism.

In Poland, the item about making cities and countrysides more beautiful is a strong indicator of a materialist attitude. In contrast to the first three clusters, the item about obtaining a stable economy and the item about progressing to a more humane society are indicative of a postmaterialist and a materialist attitude, respectively. The item regarding freedom of speech is neither an indicator of materialism nor of postmaterialism. The factor pattern of Poland is clearly different from the factor patterns of the first three clusters. As these data were collected in the beginning of the 1980s, the deviance of the pattern of Poland may be due to societal upheaval, as in those days Solidarity began to challenge the communist regime.

Results of the bottom-up procedure. Table 4 shows the results of the bottom-up procedure. The resampling method from the top-down procedure was used to determine the critical value for the pairwise similarities. As discussed, this critical value serves to determine whether the two most similar countries from the pairwise similarity matrix are similar enough to start a new phase of the bottom-up procedure. These critical values are reported in the column 'PW critical value'. The column 'BU critical value' reports the critical values from the bottom-up resampling procedure. The latter were used to determine which regions could be added to the pooled data set. The bottom-up resampling procedure was only applied at the beginning of a phase. The obtained critical value was used throughout the phase.

The two pairwise most similar regions at the beginning of a phase are shown in bold in Table 4.

TABLE 4: Results of the bottom-up procedure

Phase	Members (in alphabetical order)	PW critical value	BU critical value
1	Austria, Belgium, Canada, Denmark, France, Italy, Japan, The Netherlands, Northern Ireland, Norway, Spain, **Sweden**, **U.S.A.**, West Germany	0.9817	0.9753
2	**Chile**, Hungary, Iceland, **Ireland**, Latvia, Mexico, Moscow, Nigeria, Portugal, Russia, South Africa, United Kingdom	0.9736	0.9657
3	**Brazil,** Czechoslovakia, Estonia, **Lithuania**	0.9616	0.9466
4	**Belarus**, **Bulgaria**	0.9673	0.9527
5	**South Korea**	0.9718	--
5	**China**	0.9718	--
5	Finland	0.9718	--
5	Poland	0.9718	--
5	India	0.9718	--
5	Turkey	0.9718	--

The first set of the bottom-up procedure contains exactly the same regions as the first set of the top-down procedure. The second set also conforms quite well to the second set of the top-down approach: the former contains the same countries as the latter plus Hungary, Latvia and Moscow. This adds a bit more Eastern European 'flavour' to this second set. The third and the fourth set also mainly consist of Eastern European countries. From the fifth phase on, none of the remaining countries had a pairwise Tucker's phi that was larger than the pairwise critical value for that phase. Hence, no more equivalent sets can be formed among these countries and they have to be treated as separate cases.

The factor pattern of the first cluster is the same as was shown in Table 3 (cluster 1), that of the second cluster is not very different from the factor patterns of cluster 2 and 3 in Table 3.

Heuristic procedure. Three hierarchical clustering methods were used to reduce the dimension of the matrix of pairwise Tucker's phi coefficients between the regions, namely complete linkage, average linkage and Ward's minimum variance method (Everitt, 1993). The cluster solutions for these three methods are shown in Table 5.

All in all, there is quite some agreement between the three cluster methods. The complete linkage and the average linkage method only differed in that some clusters from the complete linkage method were split up in the average linkage method. Cluster 1 is completely the same for both methods. With a few exceptions, this also holds for Ward's method. Actually, only Nigeria and India were placed in completely different clusters than in the complete and average linkage case.

Cluster 1 from the complete and average linkage methods, and clusters 1 and 2 from Ward's method are roughly the same as the sets that were found in phase 1 of the bottom-up and the top-down procedure. Regions that are usually judged similar by the hierarchical cluster techniques, could often also be found in the same set in the bottom-up and the top-down approach.

TABLE 5: Clusters obtained via complete linkage, average linkage and Ward's minimum variance method

Cluster	Complete linkage	Average linkage	Ward's method
1	Austria, Belgium, Canada, Denmark, Finland, France, Iceland, Japan, Netherlands, Northern Ireland, Norway, Spain, Sweden, United Kingdom, USA, West Germany	Austria, Belgium, Canada, Denmark, Finland, France, Iceland, Japan, Netherlands, Northern Ireland, Norway, Spain, Sweden, United Kingdom, USA, West Germany	Canada, Finland, France, Iceland, Japan, Northern Ireland, Sweden, United Kingdom, USA
2	Brazil, Czechoslovakia, Estonia, Hungary, Latvia, Lithuania, Nigeria	Brazil, Czechoslovakia, Hungary	Austria,Belgium, Denmark, India, Netherlands, Norway, Spain, West Germany
3	Belarus, Bulgaria, Chile, Ireland, Italy, Mexico, Moscow, Portugal, Russia, South Africa	Estonia, Latvia, Lithuania, Nigeria	China, South Korea
4	Turkey	Chile, Ireland, Italy, Mexico, Portugal, South Africa	Chile, Ireland, Italy, Mexico, Nigeria, Portugal, South Africa
5	China, South Korea	Belarus, Bulgaria, Moscow, Russia	Belarus, Bulgaria, Moscow, Russia
6	Poland	India	Turkey
7	East Germany, India	East Germany	East Germany
8		China	Brazil, Czechoslovakia, Estonia, Hungary, Latvia, Lithuania
9		South Korea	Poland
10		Turkey	
11		Poland	

A more formal account of the similarity between the five procedures is given in Table 6.

TABLE 6: Proportion of identically classified countries among the five procedures

	T-D	B-U	CL	AL	W
Top-down	1				
Bottom-up	0.93	1			
Complete linkage	0.83	0.82	1		
Average linkage	0.84	0.83	0.95	1	
Ward's method	0.82	0.80	0.85	0.88	1

This table was constructed as follows. Within each procedure all countries were first compared pairwise (this yields 741 comparisons). If the countries in the comparison were in the same equivalent set or cluster for that procedure, the outcome of the comparison was 1, and 0 otherwise. Subsequently, these results were used to compare procedures. Table 6 shows, for each combination of procedures, the proportion of pairwise comparisons with the same result (a value of 1 for both procedures or a value of 0 for both procedures).

It can be seen in Table 6 that the results of the complete linkage clustering and the average linkage clustering were most similar. Also the bottom-up and the top-down procedure yield quite similar results. The similarity between the bottom-up and the top-down procedure on the one hand and the hierarchical clustering approaches on the other hand is a bit lower, but is still well above 80%. Overall, 71% of the groups involved were classified identically across all 5 procedures. In sum, the overall agreement of the procedures is substantial but far from optimal.

CONCLUSION AND DISCUSSION

The general agreement between the bottom-up and the top-down procedure on the one hand and the hierarchical clustering approaches on the other hand, to some degree supports the proposed resampling procedures for the former approaches. If a value of 0.90 as a rule of thumb were used as the critical value for Tucker's phi, the top-down procedure would yield one large set of equivalent regions, except for Poland and Lithuania. This partition of the data is clearly not supported by the results of the heuristic approach. Hence, we advise against using rules of thumb for determining critical values in the quest for equivalent partitions in a multiple group context. Chan et al. (1999) have argued similarly in the context of pairwise comparisons.

In the analysis of the World Values Survey data, Tucker's phi was used to assess the similarity between the factor loadings. It might well be, however, that other congruence indices will yield better results. Monte Carlo studies are necessary, firstly to evaluate which of the three proposed procedures yields the best results and secondly to determine whether or not some congruence indices are more suitable for use with these three procedures.

A final remark pertains to the use of Procrustes rotation in the assessment of factorial agreement. This rotation procedure has to be applied when assessing the factorial agreement of multiple constructs simultaneously, because the rotation of the factors in the latent trait space is arbitrary and hence will most certainly differ across cultural groups. The Procrustes rotation serves to rotate the factors in such a way that their agreement is maximised. It has been noted, however, that target rotation procedures are "too lenient" for the data and that rules of thumb tend to overestimate factorial similarity ((Bijnen, Van der Net and Poortinga, 1986, Van de Vijver and Poortinga, 1994). In the present illustration only one factor was extracted, in which case no rotation is possible. In our view, the resampling methods proposed here may overcome the problem of lenient criteria for evaluating factorial agreement by applying more appropriate critical values.

REFERENCES

Bijnen, E.J., T.Z. Van der Net and Y.H. & Poortinga. 1986. "On cross-cultural comparative studies with the Eysenck Personality Questionnaire." *Journal of Cross-Cultural Psychology* 17: 3-16.

Bryk, A.S., and S.W. Raudenbush. 1992. *Hierarchical linear models: Applications and data analysis.* Newbury Park: Sage.

Chan, W., R.M. Ho, K. Leung, D.K.S. Chan, and Y.F. Yung. 1999. "An alternative method for evaluating congruence coefficients with Procrustes rotation: A

bootstrap procedure." *Psychological Methods* 4: 378-402.

Efron, B. 1979. "Bootstrap methods: Another look at the jackknife." *Annals of statistics* 7: 1-26.

Espe, H. 1985. "A cross-cultural investigation of the graphic differential." *Journal of Psycholinguistic Research* 14: 97-111.

Everitt, B.S. 1993. *Cluster analysis. Third edition.* London: Edward Arnold.

Everitt, B.S. and S. Rabe-Hesketh. 1997. *The analysis of proximity data* (Kendall's library of statistics 4). London: Arnold.

Fontaine, J. 1999. *Culturele vertekening in Schwartz' waardeninstrument: een exemplarisch onderzoek naar culturele vertekening in sociaal- psychologische en persoonlijkheidsvragenlijsten*. Leuven: KUL, Faculteit Psychologische en Pedagogische wetenschappen. Unpublished Doctoral thesis.

Goldstein, H. 1987. *Multilevel models in educational and social research.* London: Griffin.

Hofstede, G. 1980. *Culture's consequences. International differences in work-related values.* Beverly Hills, CA: Sage.

Inglehart, R. 1993. *World Values Survey 1990-1991. WVS program.* J.D. Systems, S.L. ASEP S.A.

Inglehart, R. 1997. *Modernization and postmodernization. Cultural, economic and political change in 43 societies.* Princeton: Princeton University Press.

Johnson, T. 1998. "Approaches to equivalence in cross-cultural and cross-national surveys." *ZUMA Nachrichten Spezial No. 3: Cross-cultural survey equivalence*: 1-40.

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

Mooney, C.Z., and R.D. Duvall. 1993. *Bootstrapping. A nonparametric approach to statistical inference* (Sage university papers series on quantitative applications in the social sciences, series no. 07-095). Newbury Park: Sage

Muthén, B. O. 1991. "Multilevel factor analysis of class and student achievement components." *Journal of Educational Measurement* 28: 338-354.

Muthén, B. O. 1994. Multilevel covariance structure analysis. *Sociological Methods & Research* 22: 376-398.

Rensvold, R.B., and G.W. Cheung. 1998. "Testing measurement models for factorial invariance: a systematic approach." *Educational and psychological measurement* 58: 1017-1034.

Schwartz, S. H. 1992. "Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries." (Vol. 25) In *Advances in experimental social psychology*, edited by M. Zanna. New York: Academic Press.

Steenkamp, J.-B. E.M., and H. Baumgartner. 1998. "Assessing measurement invariance in cross-national consumer research." *Journal of consumer research* 25: 78-90.

Tucker, L.R. 1951. *A method for synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.

Van de Vijver, F. 1998. "Towards a theory of bias and equivalence." *ZUMA Nachrichten Spezial No. 3: Cross-cultural survey equivalence*: 41-65.

Van de Vijver, F., and K. Leung. 1997a. *Methods and Data Analysis for Cross-Cultural Research*. Thousand Oaks: Sage.

Van de Vijver, F., and K. Leung. 1997b. "Methods and data analysis of comparative research." pp. 257-300 in *Handbook of cross-cultural psychology (2nd ed., vol. 1)*, edited by J.W. Berry, Y.H. Poortinga, and J. Pandey. Boston: Allyn & Bacon.

Van de Vijver, F. & Y. Poortinga. 1994. "Methodological issues in cross-cultural studies on parental rearing behavior and psychopathology" pp. 173-197 in *Parental rearing and psychopathology*, edited by C. Perris, W.A. Arrindell, and M. Eisemann. Chicester: Wiley.

Van de Vijver, F., and Poortinga, Y. (in review). *Structural Equivalence in Multilevel Research*. Paper submitted for publication.

Welkenhuysen-Gybels, J., and J. Billiet. 1999. Een evaluatie van de crossculturele equivalentie van meetschalen rekening houdend met methode-effecten. *Mens & Maatschappij* 74: 380-391.

Welkenhuysen-Gybels, J., I. Hajnal, and J. Billiet. 2000. "On the evaluation of construct equivalence in a multigroup setting." *Paper presented at the 22nd Biennial Conference of the Society for Multivariate Analysis in the Behavioural Sciences, July 15-17 2000, London, UK.*

NOTES

1. Available online at http:/www.kuleuven.ac.be/facdep/social/soc/software.htm