The effects of measurement error in cross cultural research

Saris, Willem E.

Empfohlene Zitierung / Suggested Citation:
Saris, W. E. (1998). The effects of measurement error in cross cultural research. In J. Harkness (Ed.), *Cross-cultural survey equivalence* (pp. 67-84). Mannheim: Zentrum für Umfragen, Methoden und Analysen -ZUMA-. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-49732-6

The Effects of Measurement Error in Cross Cultural Research

WILLEM E. SARIS

In survey research many decisions are made in order to design an instrument for data collection. These choices have to do with the formulation of the question, the response categories, the instruction, the sample, the mode of data collection, etc. Each of these choices can lead to different errors (Sudman and Bradburn, 1974; Belson, 1981; Schuman and Presser, 1981; Dijkstra and Van der Zouwen, 1982; Andrews, 1984; Molenaar, 1986; Billiet et al., 1986; Groves, 1989; Alwin and Krosnick, 1991, and Scherpenzeel and Saris, 1997) and consequently to incomparability of results with respect to estimates of correlations and effect parameters across studies and also across countries. It is common knowledge that cross-cultural comparison can only be made if the measurement procedures are completely the same. In this study, we want to argue that this requirement is not enough. We will show that the results can also differ if the same procedures have been used because of differences in measurement errors in the different countries. We therefore propose a procedure to correct for measurement error, in order to make comparisons across countries with respect to correlations and regression coefficients. To correct for measurement error, we have chosen an approach that can be used by every researcher involved in social science research. This in particular is why we advocate this approach, even though, from a methodological point of view, more suitable approaches are available. We avoid using these methods because one purpose of this project is that we want to demonstrate a procedure for the correction of measurement error which can be used in any study, once prior methodological research is done. We begin with a discussion of the problems connected with measurement error in comparative survey research and then we describe the solution we have chosen for these problems. All examples given are based on the satisfaction studies done in the context of a methodological, comparative research project involving 13 language areas.

1. The effect of measurement error

The problem of measurement error in research is quite well known. These errors can bias the correlations between the variables in a study, and as a consequence, bias the estimates of parameters in models (see, for example, Bollen, 1989, chapter 5). In comparative research an extra complication is that the choices of the different instruments might make the results incomparable across countries. Let us give a simple example. In a cross cultural study (Saris et al., 1996), the same respondents were asked repeatedly to indicate their satisfaction with life in general (GLS), and their satisfaction with housing (SH), with their financial situation (SF) and with social contacts (SC). Each time the questions were presented, a different response scale was used. In the Dutch study used as an example here, the questions were presented first with a line-drawing scale and repeated with a 10-point scale in a first interview; in a second interview four weeks later, the questions were presented with a 100-point scale and a 5-point scale (for a more detailed description of the study design, see Scherpenzeel (1996)). It is therefore possible to compare the correlations between these four variables measured, using different scales for the same respondents. In Table 1 the correlation for the 1,599 respondents are presented. The coefficients of the 5-point scale and 10-point scale measures,[1] presented in Table 1 are polychoric correlation coefficients resulting from calculations with PRELIS 1 (Jöreskog and Sörbom, 1988). In the same way, data were collected in Hungary from a sample of 300 people. Here, however, three instead of four different procedures were used (Münnich, 1996). The correlations estimated in the same way for this study are presented in Table 2.

When it is realised that in each of these tables with correlation matrices, the relationships between the same variables for the same respondents are given, then it is surprising that such large differences in correlations are found between the matrices. One might think that this is related to the different points in time of some of the measures; but even when the time is held constant for the Dutch correlations, comparing the 10-point scale correlations with the line production correlations, and comparing the 5-point scale correlations with the 100-point scale correlations, the differences are still considerable. The correlations of SC with SH and SF in the 100-point matrix, for example, are twice as high as they are in the 5-point matrix, even though they were collected at the same point in time. The Hungarian correlation matrices vary just as much, but these data were all collected in one interview with the same people.

[1] Because some of the measures are categorical in nature, polychoric correlation coefficients were calculated with PRELIS 1 (Jöreskog and Sörbom, 1988) to avoid effects of categorisation of, in principle, continuous variables. The advantage of this type of coefficient is that it provides an estimate of the correlation between the variables correcting for the categorical nature of the observed variables. A categorical measure is defined as a measure with less than 15 categories used. The 100-point measures were treated as continuous when at least 15 numbers were used by the respondents. The graphical line-drawing scale was always continuous.

Table 1. Correlations between four satisfaction variables measured with four different methods obtained from the same respondents at two different points in time in the Netherlands.

				TIME 1				
	GLS	SH	SF	SC	GLS	SH	SF	SC
line production				10-point scale (polychoric corr)				
GLS	1.00				1.00			
SH	.356	1.00			.458	1.00		
SF	.370	.364	1.00		.456	.434	1.00	
SC	.454	.253	.303	1.00	.491	.325	.333	1.00
				TIME 2				
	GLS	SH	SF	SC	GLS	SH	SF	SC
100-point scale				5-point scale (polychoric corr)				
GLS	1.00				1.00			
SH	.570	1.00			.381	1.00		
SF	.544	.529	1.00		.445	.349	1.00	
SC	.644	.515	.518	1.00	.462	.232	.270	1.00

Table 2. The same data collected in Hungary.

	GLS	SH	SF	SC	GLS	SH	SF	SC
10-point scale (polychoric corr)				5-point scale (polychoric corr)				
GLS	1.00				1.00			
SH	.490	1.00			.341	1.00		
SF	.637	.468	1.00		.664	.380	1.00	
SC	.519	.254	.308	1.00	.296	.182	.247	1.00
	GLS	SH	SF	SC				
100-point scale								
GLS	1.00							
SH	.450	1.00						
SF	.614	.460	1.00					
SC	.401	.320	.310	1.00				

These differences between the different methods are clear illustrations of the problem of measurement error in survey research and we do not know what the correct estimates of the correlations between the satisfaction variables are. Since the correlations should be the same, because they represent the correlations between the same variables for the same people, the only explanation for the differences is that the methods produce different error structures, and that these errors have large effects on the correlations and consequently on all the estimates which are derived from these correlations. In this study, these questions were asked several times with different methods, allowing us to see that such differences exist. In studies where only one method is used, this cannot be seen, but the obtained correlations can be just as incorrect, because they, too, are affected by the typical errors of the specific method used.

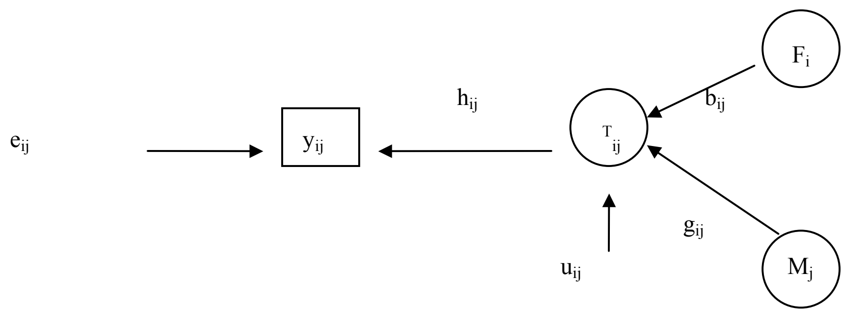
In addition, it is clear that comparisons of correlations across different countries is also very difficult, since even correlations obtained with the same measurement procedures lead to different conclusions. For example, comparing the correlation between GLS and SH for a 10-point scale would lead to the conclusion that the correlation in Hungary is higher. However, looking at the same correlation for the 100-point scale, the Dutch correlation is higher. Many similar examples can be given. These results suggest that even using the same procedure, the conclusions depend on the method which is used. It illustrates that the commonly accepted wisdom that one can only make cross-cultural comparisons if the methods are exactly the same is not, in fact, correct. The equality of the methods is neither a necessary nor sufficient condition for cross-cultural comparability. The reason for this will be clarified in the next section.

2. Explanation of the differences in correlations

Several studies have been published about measurement error and method effects (e.g., Sudman and Bradburn, 1974; Belson, 1981; Schuman and Presser, 1981; Dijkstra and Van der Zouwen, 1982; Andrews, 1984; Molenaar, 1986; Billiet et al., 1986; and Alwin and Krosnick, 1991). The approach suggested by Andrews for estimating the size of the effects of the errors and the procedure to correct for them is discussed in this paper. We have chosen this approach because it is the most explicit and general one of the different procedures introduced by these researchers. It provides all researchers, after a specialised methodological study, with information to make different measurement instruments comparable within a study and across studies. To be able to describe this approach, we first have to formulate the problem of measurement error in a more formal way. For this we use the formulation given in a publication of Saris and Andrews (1991) and Saris and Münnich (1995). In these studies, the authors suggest the path model presented in Figure 1 as a summary of their idea.

Figure 1.

A model for the response on a question incorporating method effects, unique components, and random error.



In a more formal way this idea can be formulated as follows: The responses y_{ij} on item i using method j, can be decomposed into a stable component T_{ij}, which is called the "true score" in classical test theory (Heise and Bohrnstedt, 1970; Lord and Novick, 1968) and a random error component e_{ij}. If the response variable and the variable representing the stable part are standardised, we get equation (1):

$$y_{ij} = h_{ij}\, T_{ij} + e_{ij} \qquad (1)$$

where h_{ij} represents the strength of the relationship between the stable component, or true score, and the response. The true score can further be decomposed into a component representing the score on the variable of interest, F_i, a component due to the method used, M_j, and a unique component due to the combination of method and trait, u_{ij}. After standardisation, this leads to the formulation of equation (2):

$$T_{ij} = b_{ij}F_i + g_{ij}M_j + u_{ij} \qquad (2)$$

where b_{ij} represents the strength of the relationship between the latent variable of interest and the true score and g_{ij} indicates the method effect on the true score. All variables are standardised, except for the disturbance variables. Furthermore, we assume, as is normally done, that the correlations between the disturbance variables and the explanatory variables in each equation and across equations is zero, and we assume that the method and trait factors are uncorrelated.

If all variables except the disturbance terms are standardised, the coefficients h_{ij}, b_{ij} and g_{ij} indicate the strength of the relationships between the variables in the model, and these coefficients have been given a special interpretation:

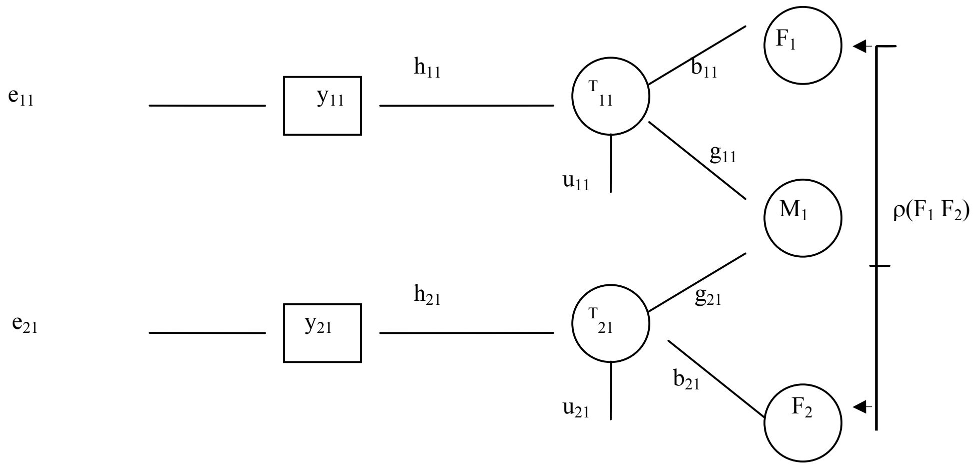
- h_{ij} is called the "reliability coefficient". The square of this coefficient is an estimate of the test-retest reliability in the sense of classical test theory (Heise and Bohrnstedt, 1970; Lord and Novick, 1968).
- b_{ij} is called the "true score validity coefficient" because the square of this coefficient is the explained variance in the true score due to the variable of interest.
- g_{ij} is called the "method effect" because the square of this coefficient is the explained variance in the true score due to the method used.
- The variance of u_{ij} plus g_{ij}^2 is sometimes called the "invalidity", because it is the variance explained in the true score which is not due to the variable of interest (Heise and Bohrnstedt, 1970).

It can be seen that with this information, the total measurement error in the responses (y_{ij}) can be decomposed into a random component (var(e_{ij})) and a systematic component (var(u_{ij}) + g_{ij}^2).
With this notation and simple path analysis, we can demonstrate all possible effects of measurement error on the correlations and effect parameters.

Figure 2 illustrates the effect of measurement error on the correlations. The only difference between Figure 1 and Figure 2 is that now two variables are studied at the same time and that we assume that these two variables are correlated. This correlation is denoted by $\rho(F_1\ F_2)$. It is assumed that the same method is used for each variable but that the method factor is uncorrelated with the trait factors. The disturbance variables are assumed to be uncorrelated with each other and the factors. All other assumptions made for the model in Figure 1 also hold, and the parameters have the same meaning as described before.

Figure 2.

A model for two correlated variables incorporating method effects, unique components, and random error.



Path analysis suggests that the correlation between the observed variables, denoted by $r(y_{11}y_{21})$ is equal to the correlation produced by F_1 and F_2 and the spurious relationship due to the method-specific variation in the observed variables. This result is specified in (3):

$$r(y_{11}y_{21}) = h_{11}*b_{11}*\rho(F_1\ F_2)*\ b_{21}*h_{21} + h_{11}*g_{11}*g_{21}*h_{21} \qquad (3)$$

Since the validity coefficients and the reliability coefficients are maximally 1, it follows from (4) that $r(y_{11}y_{21}) = \rho(F_1\ F_2)$ only if the reliability and validity are maximal and the method effect is zero. A situation like this is extremely unlikely. Therefore, the two correlations will in general be different. Since the effects of reliability, validity, and method differ from method to method, this might be the explanation for the differences in correlations found between the different methods in Table 1 and 2. The reader can easily check for him/her self that any correlation between the factors of interest can produce very different correlations, depending on the size of the validity and reliability coefficients and method effects. This variation makes it impossible to compare correlations obtained in different studies.

3. An empirical illustration

The International Research group for Methodological and Comparative Survey research (IRMCS) has done a number of projects to estimate these quality indicators for survey instruments in general. A description of the approach can be found in Saris and Münnich (1995) and Scherpenzeel and Saris (1997). An application of the approach on life satisfaction research can be found in Saris et al. (1996). For this project, in each language area, a study was carried out to obtain estimates of reliability, validity, and method effects for that country. After that, a meta-analysis was made in order to study the effects of the different characteristics of the instruments used on the validity and reliability of the instruments (for details, see Scherpenzeel, 1995). In Table 3, the results of the Scherpenzeel (1995) study are summarised.

In the first row of this table, the overall mean validity and reliability coefficients for satisfaction measures can be found. In the other rows of the table, the adjustments for this expected value are specified for different data collection situations. In each row, the adjustment for a different specific study characteristic is mentioned (for a fuller description of the table, see Scherpenzeel (1995; 64-68). It can be seen that a large variety of characteristics has been taken into account, such as the specific trait studied, the scale, the method of data collection, the position in the questionnaire, and some factors which have to do with the design of the study, such as whether an instrument is used alone or in combination with others, what the position of the instrument was in the sequence of methods used and the country in which the data collection took place.

Table 3. Meta-analysis of life satisfaction data across countries.

		Validity Coefficient Mean = .940	Reliability Coefficient Mean = .911
	N measures	Multivariate Deviations	Multivariate Deviations
SATISFACTION DOMAIN			
Life in general	54	-.006	-.038
House	54	.005	.029
Finances	54	.003	.020
Social contacts	54	-.001	-.011
RESPONSE SCALE			
100 p. number scale	64	-.021	-.027
10 p. number scale	72	.011	.051
5/4 p. category cale	72	-.022	-.026
graphical line scale	8	.058	-.007
DATA COLLECTION			
Face-to-face interview	96	.011	.012
Telephone interview	52	.002	-.051
Mail questionnaire	40	-.014	-.011
Tele-interview	28	-.022	.067
POSITION			
1 - 548		.011	.026
6 - 45	68	.017	-.001
50+ 100		-.017	-.012
TIME BETWEEN REPETITIONS			
alone in interview	32	.010	-.071
first/last 5-20 minutes	64	.017	.063
first/last 30- 60 minutes	80	-.021	-.023
middle, 5-20 minutes	16	.043	.028
middle, 30-60 minutes	24	-.017	-.016
ORDER OF PRESENTATION			
first measurement	60	-.015	-.025
repetition	156	.006	.010
COUNTRY			
Slovenia	12	.020	-.013
Germany	16	.007	.028
Catalonia (Spain)	12	-.039	-.022
Italy 12		.013	.043
Flanders (Belg)+ Netherlands	64	-.028	-.039
Wallonia (Belgium)	12	-.026	-.028
Brussels (Belgium)	12	.006	.000
Sweden	12	.023	.099
Hungary	12	.050	.046
Norway	16	-.018	.031
Russians (Russia)	12	.043	.004
Tatarians (Russia)	12	.033	.003
Other nationalities in Russia	12	.039	.000

The last point is of special interest to us. This study suggests, for example, that on average the validity will be .94, but depending on the chosen instruments, this quality indicator will be higher or lower.

Table 4. Prediction of the validity and reliability of a measure in the Dutch study, on the basis of the instrument characteristics.

	Validity coefficient	Reliability coefficient
	Mean = .940	Mean = .911
Adjustments for:		
Domain: GLS	-.006	-.038
10-point scale	+.011	+.051
Data collection by mail	-.014	-.011
Position 6-45	+.017	-.001
Design time: alone	+.010	-.071
Design order: first	-.015	-.025
The Netherlands	-.028	-.039
Sum	.915	.777

Table 5. Prediction of the validity and reliability of a measure in a Hungarian study, on the basis of the instrument characteristics.

	Validity coefficient	Reliability coefficient
	Mean = .940	Mean = .911
Adjustments for:		
Domain: GLS	-.006	-.038
10-point scale	+.011	+.051
Data collection by mail	-.014	-.011
Position 6-45	+.017	-.001
Design time: alone	+.010	-.071
Design order: first	-.015	-.025
Hungary	+.050	+.046
Sum	.993	.862

On the other hand, even if the instruments are identical in two countries, the validity can be different due to country-specific differences. For example, the validity in Slovenia will on average be .02 higher than the mean, while in Catalonia the validity on average will be .039 lower. Similar effects can be found for other countries and for reliability. This suggests that the quality of the data differs from country to country, even if they use the same data collection procedure. We illustrate this important point below. Any researcher who has one measure of a satisfaction variable can determine the quality of this measure on the basis of the results presented in Table 3. For example, if we say GLS was measured by mail using a 10-point scale at the beginning of the interview in the Netherlands and in Hungary, we can estimate the validity and reliability coefficients with the information from Table 3, as shown in Table 4 and 5. By adding up all the adjustments to the mean value, we obtain an estimate of the validity and reliability coefficient for this variable. For the Dutch study the result is presented in Table 4, for Hungary, in Table 5.

The tables indicate that even if the same instruments are used in both countries for measurement of satisfaction, large differences in results are found for the Netherlands and Hungary. In the same way, these two coefficients can be estimated for the other traits and other methods. For the Dutch and Hungarian study, the results of these calculations for all satisfaction traits using the same method (10-point scale) are presented in Table 6.

Table 6. Quality estimates of the indicators in the MTMM study, predicted on the basis of the meta-analysis for the Netherlands and Hungary.

		Validity		Reliability		Method Effect	
		NL	H	NL	H	NL	H
10-point scale							
	GLS	.92	.99	.78	.86	.39	.14
	SH	.93	1.0	.85	.93	.37	.00
	SF	.93	1.0	.84	.92	.37	.00
	SC	.92	.99	.81	.89	.39	.14

In Table 6, the method effects are also included. This effect can easily be calculated from the information on the validity coefficient, because the method variance should be 1- b^2_{ij} if the unique variance is zero[2]. So the estimate of the method effect parameter is the square root of the method variance, or:

$$g_{ij} = \sqrt{(1- b^2_{ij})} \qquad (4)$$

If the measurement procedure indicated above is used, in both countries the reliability, validity and method effects for both variables will be different, as demonstrated above. Using equation 3, it can be shown that in that case the correlation will also be different, although the correlation was the same between the variables of interest. For instance, if we take a correlation of .8 for the variables GLS and SH, for the Netherlands we would get:

$$r\,(R1,R2) = .78*.92*(.8)*.93*.85 + .78*.39*.37*.85 = .45 + .096 = .55$$

In the same way, for Hungary we would get:

$$r\,(R1, R2) = .86*.99*(.8)*1.0*.93 + .86*.14*.00*.93 = .63 + .00 = .63$$

First of all we see that the resulting correlation is much lower in both cases than the correlation between the variables of interest due to the relatively low reliability. In addition, we see a difference of .08 between the resulting correlation in the two countries, even though the correlations between the theoretical variables in both countries were identical. This difference has no substantive meaning, it is only due to the difference in quality of the measurement procedures in the two countries. It seems that the Hungarian public, somehow less bothered by questionnaires, gives better answers to the same questions than Dutch respondents do.

This result indicates that comparisons between correlations from different countries cannot be made without correction for measurement error. How these corrections can be made is the subject of the next section.

[2] This assumption is necessary for identification of the model. This assumption is realistic if in the experiment exactly the same question is used combined with each method. For details we refer to Saris (1990).

4. Correction for measurement error

Now we will concentrate on the correction for the effect of the specific method on the obtained correlation. In other words, we are interested in the correlation between the latent factors, and not in the correlations between the observed variables. To derive these correlations, we have to express the correlation between the factors in the observed correlations and the different validity's, reliability's and method effects. This expression follows immediately from equation (3):

$$\rho(F_1 F_2) = [r(y_{11}y_{21}) - (h_{11}*g_{11}*g_{21}*h_{21})] / (h_{11}*b_{11}*b_{21}*h_{21}) \qquad (5)$$

This result suggests that the correlation between the factors can be estimated simply from the observed correlation if estimates for the validity and reliability coefficients and the method effects are known. Table 3 above provides the information from which the reliability, validity and method effects for different measurement instruments can be derived. These results can be used, as before, to estimate the correlation between the variables of interest corrected for measurement error. This could be done by hand, but it is also possible to use programs like LISREL (Jöreskog and Sörbom, 1989) to estimate the corrected correlations, using the model specified in Figure 2, or a larger model for all traits for which data have been collected. Appendix A provides the LISREL input for such an analysis.

Below we give some examples using equation 5 or Figure 2. First of all, the example of the last section can be reversed. For the instruments presented in Table 6, the validity, reliability and method effects were calculated. If in the Netherlands a correlation of .55 is obtained with these instruments, and in Hungary a correlation of .63, then equation 5 can be used to show that in both countries the correlation between the two variables, corrected for measure-ment error, is identical and equal to .8.

On the other hand, if, under these conditions in both countries, a correlation between GLS and SH of .63 is found, then, using equation 5 and the results of Table 3, it can be shown that the correlation between these variables, corrected for measurements error is .95 in the Netherlands, and .80 in Hungary.

This example shows that equal correlations obtained with identical instruments can be due to quite different correlations between the variables of interest. This means that by using this correction for measurement, one can control for differences in error structures between countries and make the results comparable.

5. Conclusion

In all textbooks about structural equation models, a multiple indicators approach is recommended for the estimation of, and correction for, measurement error. Although this approach is statistically correct, many practical and substantive problems are associated with it.

First of all, it is rather expensive to measure each theoretical variable in at least two different ways. It means that one doubles the interview time, which usually is quite costly.

Second, it is difficult to ask the same question twice in one interview. Although possible, it is not easy to organise, and one risks irritating respon-dents who notice the repetition. As a substitute, researchers often vary the formulation of the repeated question. However, Heise (1969) and Saris (1982) have argued that variation in question wording might change the meaning of the variable one measures. There are, moreover, many studies which demonstrate this point, even for the mean and variance of the variables (see Schuman and Presser (1981); Belson (1981)). Consequently, it is not clear what a multiple indicator model in such a situation represents. The latent variable will be a common factor of two or more indicators, but because these indicators are substantively different, it is unclear what this common factor stands for.

On the other hand, correction for measurement error seems to be a necessity, as we have tried to indicate. We have shown that the commonly accepted idea that results can only be compared across countries if the same method has been used is, in fact, incorrect. Even if the same method is used, one can get different results due to differences in the error structure in the different countries. Therefore, correction for measurement error is necessary. Corrected correlation coefficients are more comparable, not only across different studies but also across different countries. Also, the correction for measurement error provides a better estimate of the explained variance in each equation. This is important for the evaluation of the quality of different explanatory models.

We hope to have indicated in this chapter that the proposed procedure allows correction for measurement error even if only one indicator is used for each theoretical variable. When large methodological studies as described in Scherpenzeel and Saris (1997) are involved, and tables like Table 3 here are constructed for more topics than life satisfaction (see, for example, Andrews, 1984; Rodgers et al., 1992; Költringer, 1993; Scherpenzeel, 1995), the procedure described here can be used for any correlation matrix and any structural equation model. This is what makes it an attractive approach for national and cross-national studies.

The discussion in this paper has been limited to the effect of measurement error on the correlation between variables in cross-cultural research. There are, of course, more reasons for incomparability, such as coverage differences and fieldwork differences, mode effects, etc. The discussion has focused on problems with respect to the correlations; one can also study the effect on distributions of variables. A more general approach, covering a wider range of issues, can be found in Saris and Kaase (1997). Here we have concentrated on the misleading assumption that equality of measurement procedures is sufficient to guarantee comparability in cross-cultural research. We have shown that the situation is much more complex. Without correction for measurement error in each separate study, comparability is not guaranteed. We have also shown that many methodological studies are available to realise these corrections for measurement error.

References

Alwin, D. F. & Krosnick, J. A. (1991): The reliability of survey attitude measurement: the influence of question and respondent attributes. *Sociological Methods and Research 20*: 139-181.

Andrews, F. M. (1984): Construct validity and error components of survey measures: a structural modelling approach. *Public Opinion Quarterly 48*: 409-422.

Belson, W. (1981): *The design and understanding of survey questions*. London: Gower.

Billiet, J., Loosveldt, G. & Waterplas, L. (1986): *Het survey-interview onderzocht: effecten van het ontwerp en gebruik van vragenlijsten op de kwaliteit van de antwoorden*. Leuven: Sociologisch Onderzoeksinstituut KU Leuven.

Bollen, K. A. (1989): *Structural equations with latent variables.* New York: Wiley.

Dijkstra, W. & Zouwen, J., van der (1982): *Response Behaviour in the Survey-Interview*. London: Academic Press.

Groves, R. M. (1989): *Survey Errors and Survey Costs*. New York: Wiley and Sons.

Heise, D. R. (1969): Separating reliability and stability in test-retest correlation. *American Sociological Review 34:* 93-101.

Heise, D. R. & Bohrnstedt, G. W. (1970): Validity, invalidity, and reliabiity. In: Borgatta, E. F. & Bohrnstedt, G. W. (eds.). *Sociological Methodology*. San Francisco: Jossey-Bass.

Jöreskog, K. G. & Sörbom, D. (1988): *Prelis: a program for multivariate data screening and data summarization* (second edition). Mooresville: Scientific Software.

Jöreskog, K. G. & Sörbom, D. (1989): *Lisrel VII: Users reference guide*. Mooresville: Scientific Software.

Költringer, R. (1993): *Messqualität in der sozialwissenschaftlichen Umfrageforschung*. Wien: Endbericht Project P8690-SOZ des Fonds zur Förderung der wissenschaftlichen Forschung (FWF).

Lord, F. & Novick, M. R. (1968): *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Molenaar, N. J. (1986): *Formuleringseffecten in Survey-interviews*. Amsterdam: VU-uitgeverij.

Rodgers, W. L., Andrews, F. M. & Herzog, A. R. (1992): Quality of Survey Measures: a structural modeling approach. *Journal of Official Statistics 8:* 251-275.

Saris, W. E. (1982): Different questions, different variables. In: C. Fornell (eds*.). A second generation of multivariate analysis: Vol. 2. Measurement and Evaluation*. New York: Praeger.

Saris, W. E. (1990): The choice of a model for evaluation of measurement instruments. In: Saris, W.E. & Meurs, A., van (eds.). *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies*. Amsterdam: North Holland.

Saris, W. E. & Andrews, F. M. (1991): Evaluation of measurement instruments using a structural modelling approach. In: Biemer, P. P., Groves, R.M., Lyberg, L.E., Mathiowetz, N. and Sudman, S. (eds.). *Measurement Errors in Surveys*. New York: Wiley and Sons.

Saris, W. E. & Münnich A. (1995): *The Multitrait-Multimethod approach to evaluate measurement instruments.* Budapest: Eötvös University Press.

Saris, W. E., Veenhoven, R., Scherpenzeel, A. & Bunting, B. (eds.) (1996): *Life Satisfaction in West and Eastern Europe*. Budapest: Eötvös University Press.

Saris, W. E. & Kaase M. (eds.) (1997). *Eurobarometer: measurement instrument for opinions in Europe* ZUMA-Nachrichten Spezial No. 2. Mannheim: ZUMA.

Scherpenzeel, A. C. (1995): *A Question of Quality: Evaluating survey questions by MTMM studies.* Ph.D thesis. Amsterdam: University of Amsterdam.

Scherpenzeel, A. (1996): Life Satisfaction in the Netherlands. In: Saris, W. E., Veenhoven, R, Scherpenzeel, A. & Bunting, B. (eds.): *Life Satisfaction in West and Eastern Europe.* Budapest: Eötvös University Press. Chapter 3.

Scherpenzeel, A. & Saris, W. E. (1997): The validity and reliability of survey questions: A Meta analysis of MTMM studies, *Sociological Methods and Research 25*, 341-383.

Schuman, H. & Presser, S. (1981): *Questions and answers in attitude surveys: experiments on question form, wording and context*. New York: Academic Press.

Sudman, S. & Bradburn, N. L. (1974): *Response Effects in Surveys.* Chicago: Aldin.

Appendix A.

LISREL input to estimate corrected correlations between four satisfaction variables.

```
        Satisfaction Netherlands, 5p scales, correction on basis of meta-analysis
da ni=4 no=1599 ma=pm
la
*
'sat5p1'  'sat5p2'  'sat5p3'  'sat5p4'
pm file=sat5p.pm
model ny=4 nk=5 ne=4 te=fi ga=fi ps=ze ph=sy,fr
le
*
'truesco1' 'truesco2' 'truesco3' 'truesco4'
lk
*
'general' 'house' 'financial' 'contacts' '5p'
value .86 ly 1 1
value .93 ly 2 2
value .92 ly 3 3
value .89 ly 4 4
value .89 ga 1 1
value .90 ga 2 2 ga 3 3
value .89 ga 4 4
value .46 ga 1 5
value .44 ga 2 5 ga 3 5
value .45 ga 4 5
value .26 te 1 1
value .14 te 2 2
value .16 te 3 3
value .21 te 4 4
fi ph 4 5 ph 3 5 ph 2 5 ph 1 5
fi ph 1 1 ph 2 2 ph 3 3 ph 4 4 ph 5 5
value 1 ph 1 1 ph 2 2 ph 3 3 ph 4 4 ph 5 5
start .5 all
output ns ss
```