chapter 5

The choice of a model for evaluation of measurement instruments

W.E. SARIS

In the previous chapters the problems which occur when the MTMM approach is used in the evaluation of measurement instruments have been discussed. These problems led to a lengthy discussion of the advantages and disadvantages of the different models and their necessary restrictions. In this chapter we report on the results of this discussion. We first discuss the differences between the two approaches: the MTMM and the RMM approach. Memory effects and the technical problems mentioned earlier are then discussed. Finally a design for evaluation studies will be proposed.

DIFFERENCES BETWEEN THE MTMM AND RMM MODEL

In order to clarify the differences between the different models we will start with the presentation of a measurement model which contains the characteristics common to both models (figure 1). This is the model used by Alwin and Jackson (1979) and by Heise and Bohrnstedt (1970).

In this model a distinction is made between theoretical variables that we are interested in (F_i) but which are not observable, and observed variables (x_{ij}) and true scores (T_{ij}). The latter variables are equal to the observed variables corrected for random measurement error. Furthermore, latent variables M_1 and M_2 are introduced to indicate the covariance between the true scores, which is due to the specific methods used to measure the different variables. Finally, unique variance (u_{ij}), which is specific for specific true scores, is added to allow for the possibility that the true scores for the same factor differ not only by the variance due to the common method but also because of an interaction between the method and the specific item.

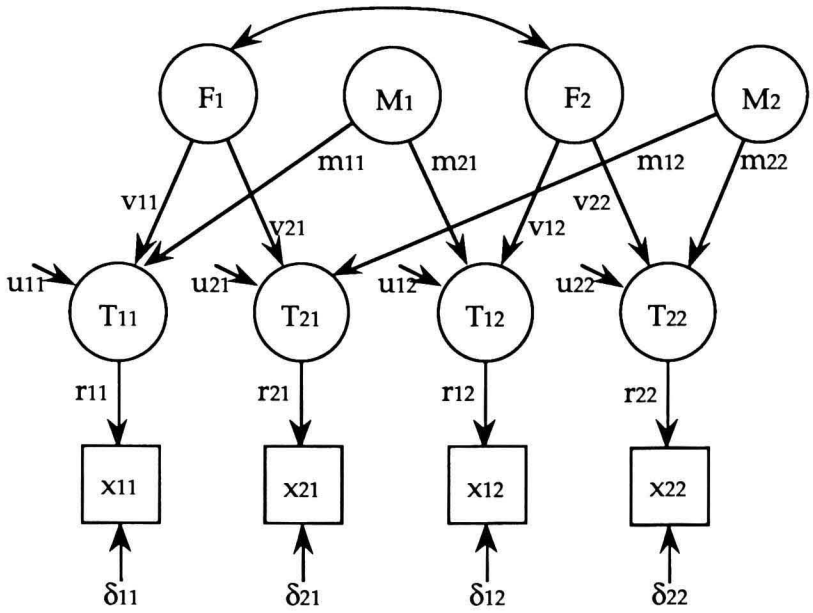

figure 1: A general measurement model

Following the definitions suggested by Saris in this volume, v_{ij} represents the validity of the jth variable measured with the ith method; $1-v_{ij}^2$ represents the invalidity of variable j measured with the ith method. A part of this invalidity is method specific variance which can be decomposed in method specific variance m_{ij}^2, where m_{ij} represents the method effect of method i on variable j, and the unique variance, $var(u_{ij})$, for the same variable. Finally, r_{ij}^2 is the reliability of variable j measured with the ith method.

Although these coefficients are not identified in this model, their meaning is clear. For their estimation at least two observations with each instrument are necessary and at least three different methods should be used for each variable. However, such an experiment is unpractical but the model is nevertheless useful to clarify the meaning of the different parameters.

If the RMM approach is used all the different coefficients which have been specified in the figure, except the method effects (m_{ij}), can be estimated. The estimation of these coefficients would require that at least three theoretical variables are measured in three different ways. We would then have to use a repeated MTMM model, which is unpractical because of the frequent repetition of the different questions (6 times for each different topic). However,

if one of the designs which has been specified by Saris is used, estimates of all other parameters can be obtained as well as very detailed information about the size of the total systematic error, the validity and the reliability as defined above. An estimate of the systematic error or unique variance in the different measurement instruments is also obtained, which is especially attractive. This allows the possibility of determining whether the different methods really measure the same variable or not.

Although the information obtained by the RMM approach seems attractive as it gives the information which we need to evaluate measurement instruments one can also argue in favour of the coefficients obtained from the MTMM approach. In the MTMM approach the true scores are ignored because only one observation is made with each instrument. Consequently the distinction between u_{ij} and e_{ij} cannot be made. The coefficient, which is called the "validity coefficient" (λ_{fij}), is the product of $v_{ij}*r_{ij}$, the so-called "method effect" (λ_{mij}) is equal to $m_{ij}*r_{ij}$ and the so called "reliability" (R^2) is equal to $(v_{ij}*r_{ij})^2 + (m_{ij}*r_{ij})^2$.

Normally the MTMM design does not facilitate the estimation of the same coefficients as the RMM model. Only if there is no other source of systematic error beside the method effects can the coefficients of figure 1 be derived using the MTMM approach. In that case $R^2 = r_{ij}^2$ and on the basis of this information the other coefficients can also be obtained from the estimated parameters. A disadvantage of this approach is that one can not test this condition, which is not likely to be fulfilled. This does not mean, however, that the obtained information is irrelevant.

The results obtained suggest that the MTMM approach provides us with estimates of the effects of the latent variables on the observed variables, which are called the validity (λ_{fij}), the method effect (λ_{mij}) and the explained variance of the observed variables, which are called the reliability. Although these coefficients are not pure estimates of the characteristics of the measurement instruments, these parameters are certainly useful if relationships between different variables are studied. The validity and reliability coefficients are necessary to obtain consistent estimates of the correlations and effects between the theoretical variables. For example from figure 1 we get:

$$r(x_{11},x_{12}) = \lambda_{f11} * r(F_1,F_2) * \lambda_{f12} + \lambda_{m11} * \lambda_{m12}$$

and from this result follows:

$$r(F_1,F_2) = \frac{r(x_{11},x_{12}) - \lambda_{m11} * \lambda_{m12}}{\lambda_{f11} * \lambda_{f12}}$$

Thus in order to derive the correlation between the factors of interest we need estimates of the effects which can be obtained using the MTMM approach. The argument given above suggests that

- With the MTMM approach item specific variance and measurement error cannot be distinguished, and therefore one can not say whether the questions measure the same variable except for method specific variance.
- With the RMM approach the method effects cannot be detected, and therefore the inflation in the correlations between the observed variables, which is due to method correlation, cannot be corrected.

If the researcher ensures that exactly the same question is asked and only the method is varied, the item specific variance will probably be very small and can be ignored. However, the lack of information about the method effects seems to be a serious omission.

It would seem that for theoretical reasons, in order to determine whether questions really measure the same variable, the RMM approach is preferable. On the other hand, for practical purposes the coefficients of the MTMM approach, such as validity coefficients and method effects, are very important for deriving consistent estimates of the parameters of interest: the correlations between latent variables and the effects of latent variables on each other. For this reason the MTMM approach seems preferable.

Although this seems the proper conclusion it is important to stress its consequences: one has to use the same formulation of the question for all different methods. This rule has to be taken very literally. Let us apply it to satisfaction with health and examine what it means for three different methods:

How satisfied are you with your health?
Choose an answer from the following categories

1. very satisfied,
2. satisfied,

3. neither satisfied nor dissatisfied,
4. dissatisfied,
5. very dissatisfied

How satisfied are you with your health?
give a number between 1 and 10

1= very satisfied
10= very dissatisfied

These questions are the same and only the measurement scale is different. Even in such a case there would be some doubt as to whether the questions measure the same opinion, but this assumption can be accepted without a test. For evidence on this point we refer to Saris (1982). If the following question is used a test would certainly be necessary:

Some people worry about their health a lot. How about you - how worried are you about your health?

1. very worried,
2. quite worried
3. a little worried
4. not at all worried

This question introduces a new component: the emphasis of the question has shifted from "satisfaction" to "worry", which are not the same concepts. There can be considerable doubt about the equality of these three questions. Therefore, it is not surprising that Rodgers et al. (1988) found very low validity, high method effects and very high residual variance between these questions. This could be explained by item specific variation; in other words, the three questions do not measure the same variable.

This type of variation in questions should be avoided in the experiments. If there is any doubt about the possibility of item specific variance a RMM experiment should be performed to test for it or the whole MTMM questionnaire should be repeated later in a panel to see if a specific component exists. This valuable suggestion was made by Andrews on the basis of his recent research.

Repetition of questions and memory effect

A criticism of the RMM model is that it requires the repetition of various questions twice in one interview session. This procedure seemed questionable to some participants, given the possibility of memory effects. A simple experiment with a questionnaire showed that the same problems occur in the MTMM approach. This does not alleviate the problem in any way. In all cases the possibility of memory effects must be taken into account. So far it is not clear how this has been done in previous experiments. Its necessity can be illustrated by the study of Hox (1986) who asked questions on 5 different aspects of life satisfaction with 5 different methods in one mail questionnaire. Looking at the correlation matrix obtained we see that the correlations of variables measuring the same trait with different methods increases dramatically from around .5 in the beginning of the interview to around .8 at the end. It is not clear what this means but the most likely explanation is a memory effect combined with consciousness raising effects due to the repeated measurement.

In order to check this hypothesis it was decided that an experiment should be carried out to determine how many questions are necessary between repeated questions, to counteract such an effect.

SOME TECHNICAL PROBLEMS

Before a decision can be made in favour of the MTMM approach it should be clear that the design and model do not generate too many technical problems. These problems have also been discussed and an overview of the chosen solutions will be given.

Convergence problems

The MTMM model was heavily criticized for convergence problems when the data were analyzed with LISREL. In many instances the program could not estimate the parameters of the model. It was suggested by Költringer (1990) and Saris (1990) that this was a consequence of overfactoring. These remarks were in line with earlier remarks of Rindskopf (1984). Andrews, who did not have these problems in his study, suggested that this problem could be due to the lack of restrictions in the model. In his study he had introduced constraints on the method effects, specifying that they should be the same for all traits. This seems to be a plausible

assumption. This suggestion was therefore evaluated on different data sets.

In data sets of Saris (1990) all MTMM models with the constraints mentioned above converged if the method factors were uncorrelated. For the same data the estimation did not converge in 5 out of 8 cases with the same constraints and unconstrained correlations between the method factors. Unacceptable solutions were found in the other three cases. The same problem has also been detected by Rodgers et al. (1988). When these correlations were restricted (fixing some at zero, leaving others free) mixed results were obtained: acceptable solutions were sometimes obtained, but frequently not. We will come back to the problem of correlated factors in one of the following sections.

Identification and design

Andrews also suggested that the problems caused by the repetition of questions could be reduced by using a 4 by 2 design, i.e. with four traits and two methods. This remark was however contrary to that of Költringer who held that at least three methods are required for identification. The 4 by 2 MTMM model without any constraints would not be identified. However, if the equality constraints are introduced on the method effects, as specified above, then the model would be identified. This does not imply that this 4 by 2 model should be used. In chapter 8 it has been shown that very different as well as irregular results can be obtained with different 4 by 2 models derived from a data collection using a 4 by 3 design. It was shown that the results are very unstable due to the fact that only two indicators per factor were used. This problem has also been mentioned by Boomsma (1983).

Memory effects and correlations between the method factors

During the conference Saris presented two correlation matrices for the same variables obtained from different groups, one of which had no knowledge about the topic and the other with some information. These matrices were completely different. In the matrix of the uninformed people high method effects could be seen and correlations between the substantive variables were low. In the matrix for the informed people the method effects were very small but the substantive correlations were much higher. This result had been replicated in other studies as well. In these studies correlated method effects were not necessary at all for informed people, while for the other topics about which the same respon-

dents had little knowledge, correlations between line and number responses were needed.

It seems that the respondents answer questions differently if they have an opinion than when they are uninformed. In the latter case they try to be consistent between methods, not caring about what they say. This can explain the high method effects and high correlations between the methods in the uninformed case. Given these results it seems that for the time being topics about which the respondents have little knowledge should be avoided. It could perhaps be shown that these effects are smaller if the interval between the various repetitions of the questions is longer.

THE CHOSEN PROCEDURE

The discussion reported above has led to a number of decisions with respect to the approach to be used and these will be reported below.

Model choice and model test

A test to chose a model from the possible sets of models suggested by Költringer and Saris is necessary. It was shown that the χ^2 test which is normally used leads to many difficulties because of the varying power of the test in different data sets.

Therefore the general conclusion was that testing should be avoided as much as possible and one specific model should be fitted to all data sets. As we want to estimate the validity, invalidity and random measurement error, the best candidate is the model of Andrews which assumes the method effects are identical. This factor model is specified as follows:

$$Y = \Lambda F + E$$

where
- Y is a 12-1 vector of observed variables (four traits, each measured with three different methods)
- F is a 7-1 vector containing the four trait and three method factors
- E is a 12-1 vector of random error components
- Λ is a 12-7 matrix of effect parameters

The covariance matrix of the error components (E) of this model is specified as a diagonal matrix.

table 1: Specifications of the matrix Λ

	F_1 (trait1)	F_2 (trait2)	F_3 (trait3)	F_4 (trait4)	F_5 (method1)	F_6 (method2)	F_7 (method3)
Y_1	$\lambda_{1,1}$				$\lambda_{1,5}$		
Y_2	$\lambda_{2,1}$					$\lambda_{2,6}$	
Y_3	$\lambda_{3,1}$						$\lambda_{3,7}$
Y_4		$\lambda_{4,2}$			$\lambda_{1,5}$		
Y_5		$\lambda_{5,2}$				$\lambda_{2,6}$	
Y_6		$\lambda_{6,2}$					$\lambda_{3,7}$
Y_7			$\lambda_{7,3}$		$\lambda_{1,5}$		
Y_8			$\lambda_{8,3}$			$\lambda_{2,6}$	
Y_9			$\lambda_{9,3}$				$\lambda_{3,7}$
Y_{10}				$\lambda_{10,4}$	$\lambda_{1,5}$		
Y_{11}				$\lambda_{11,4}$		$\lambda_{2,6}$	
Y_{12}				$\lambda_{12,4}$			$\lambda_{3,7}$

table 2: Specifications of the correlation matrix of the factors

	F_1 (trait1)	F_2 (trait2)	F_3 (trait3)	F_4 (trait4)	F_5 (method1)	F_6 (method2)	F_7 (method3)
F_1	1.0						
F_2	$ps_{2,1}$	1.0					
F_3	$ps_{3,1}$	$ps_{3,2}$	1.0				
F_4	$ps_{4,1}$	$ps_{4,2}$	$ps_{4,3}$	1.0			
F_5	0	0	0	0	1.0		
F_6	0	0	0	0	0	1.0	
F_7	0	0	0	0	0	0	1.0

If the respondents have an opinion on the question topic and the time between the repetitions is not too short, this model will be quite close to the true model for the MTMM experiments. Incidentally, correlations between the factors may be needed. In that case these correlations are introduced one after the other while combinations of correlations are avoided if they are not absolutely necessary. The reason for this is that the introduction of correlations has an effect on the estimates of validity and invalidity.

Therefore it should be ascertained which correlations have to be introduced and which left out.

The evaluation of the model should be done on the basis of the reduction in χ^2 (taking into account the power of the test), the size of the estimated correlation coefficient and the lack of alternative possibilities of correcting the model. In order to introduce a correlation between the methods a very large improvement in fit and a large coefficient are needed. In addition, there should be no other possible correlation between the method factors. A large correlation coefficient has a value of .4 or larger. A smaller value has no effect at all on the model and the fit.

With these rules most data sets will lead to an acceptable model. Occasionally two correlations will be needed but one should then seriously consider whether there are other problems in the data collection such as too short intervals between a repetition or lack of knowledge on the part of the respondents.

Estimation and standard errors

Satorra (1990) clearly indicated that there is no problem with respect to estimation in case the normality assumption is not satisfied. The estimated values of the parameters are always consistent. The χ^2 test and the standard errors might be wrong in that case but if we assume the independence of the errors from the factors instead of the the uncorrelatedness these statistics are correct then even if the ML estimators are used. In general, this alternative assumption is reasonable, but an exception will be indicated below. The use of the asymptotic distribution free (ADF) methods leads to more problems than solutions and should, therefore, be avoided.

There is one case which requires further consideration: the situation where categorical variables with few categories are used. Here the relationships between the factors and the observed variables become nonlinear and the errors are not independent of the factors. This case requires special attention. The general consensus was that in this case the polychoric or polyserial correlations should not be used as substitutes for the normal product moment correlations because of the assumption of normality which is once again introduced. For the time being it is perhaps the best to avoid categorical scales with fewer than 5 categories (to mention an arbitrary cutting point).

Data cleaning

In the meetings a plea was made for cleaning of the data before the analyses are done, as outliers can affect the results considerably.The following steps should be carried out:

1. Look at scattergrams or tables of different methods for each trait. If outliers are found in the corners of the table or plot, list the scores of these respondents.
2. Look at the scores. Are these mistakes systematic errors which can be corrected or not? Systematic errors are reversals of the scales or due to use of too short a scale (a scale of 100 points instead of 1000 points for example). Such errors can be corrected and the case can be saved. Incidental errors like forgetting a zero (100 instead of 1000) can also be corrected if other judgments clearly indicate that this error has occurred.
   If no correction rule can be specified than the value should be made missing.
3. Look at the scattergram and tables again to see whether the picture has improved. If not, repeat the same procedure.

When making such corrections, one should of course be very careful not to improve one method while neglecting another. The number of corrections for the different methods should be registered for later analysis.

The procedure for the data analysis has been specified on the basis of this discussion. There are two issues which will be decided later: problems with category scales with few categories, and the number of cases needed for the study. On both points further studies will be done.

REFERENCES

Alwin, D.F., & Jackson, D.J. (1979). Measurement models for response errors in surveys: Issues and applications. In K.F. Schuessler (Ed.), *Sociological methodology 1980*. San Francisco: Jossey-Bass.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality*. Amsterdam: Sociometric Research Foundation.

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

Hox, J.J.C.M. (1986). *Het gebruik van hulptheorieën bij operationalisering, een studie rond het begrip welbevinden.* Amsterdam: Universiteit van Amsterdam.

Költringer, R. (1990). Analysis of mutitrait multimethod matrices. In W.E. Saris and A. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait multimethod studies.* Amsterdam: North-Holland.

Rindskopf, D. (1984). Structural equation models: empirical identification, Heywood cases, and related problems. *Sociological Methods and Research, 13,* 109-119.

Rodgers, W.L., Herzog, A.R., & Andrews, F.M. (1988). Interviewing older adults: Validity of self-reports of satisfaction. *Psychology and Aging, 3,* 264-272.

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

Saris, W.E. (1990). Models for evaluation of measurement instruments. In W.E. Saris and A. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait multimethod studies.* Amsterdam: North-Holland.

Saris, W.E. (1990). The choice of a research design for MTMM studies. In W.E. Saris and A. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait multimethod studies.* Amsterdam: North-Holland.

Satorra, A. (1990). Robustness issues in the analysis of MTMM and RMM models. In W.E. Saris and A. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait multimethod studies.* Amsterdam: North-Holland.