Developments in Survey Methodology
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 15, Ljubljana: FDV, 2000

Systematic and Random Method Effects. Estimating Method Bias and Method Variance

Germà Coenders[1] and Willem E. Saris[2]

Abstract

Two lines of research have been followed to assess the effect of the method of measurement on data quality. Split-ballot experiments allowed researchers to assess the effect of the measurement method on the marginal distribution of the responses and to make assessments of relative bias. Structural equation models allowed researchers to assess the effect of the method on reliability and validity.

This paper develops and illustrates a strategy based on fitting mean-and-covariance structure models to multitrait-multimethod data, which allows researchers to assess relative bias, reliability and validity simultaneously. Two major advantages of this approach over split-ballot designs are that relative bias is assessed after partialling out the effects of measurement errors and that alternative definitions of relative bias are possible. A complete sequence of statistical tests of relative unbiasedness of methods is provided and applied.

1 Introduction

In the design of survey research, choices must be made with regard to the wording of questions, the response scales, the question context and the technique for data collection. Each of these choices and combinations of choices lead to different errors. There have been numerous experimental studies of the effects of variations in the characteristics of survey questions (for reviews see Billiet, Loosveldt, and Waterplas, 1986; Dijkstra and van der Zouwen, 1982; Groves, 1989; Schuman and Presser, 1981; or Sudman and Bradburn, 1982). These studies illustrate that

[1] Department of Economics, University of Girona. Faculty of Economics, Campus of Montilivi, 17071 Girona, Spain. The author was supported by the University of Girona Grant S-UdG97-178.

[2] Department of Methods and Techniques. Faculty of Political Sciences. University of Amsterdam. O.Z. Achterburgwal 237. 1012 DL Amsterdam, The Netherlands.

differences in the response distributions may be obtained depending on the procedure used, but these studies have not resulted in general rules connecting the degree of measurement error with different combinations of survey and question characteristics. On the contrary, they assess only relative bias, and they usually do so by using *split ballot* experimental designs and varying one single characteristic of the measurement instrument at a time. An exception is the study of Molenaar (1986) who evaluated the simultaneous effects of a large number of question characteristics of a random sample of survey questions but estimated the effects on the frequency distribution of the observed variables, thus concentrating on *relative bias* and letting aside other aspects of data quality, such as *validity* and *reliability*.

Andrews (1984) was the first to conduct a systematic and comprehensive study of validity and reliability by meta-analyzing the estimates obtained from a large number multitrait-multimethod (MTMM) studies carried out in the USA. Under this approach, estimates of reliability and validity are obtained in a first stage for a large set of measurement instruments by fitting some form of factor analysis model to a series of data sets collected with a MTMM design. In a second stage, the variation in the data quality estimates is explained by the variation in the characteristics of the survey questions. In this way, reliability and validity of survey measurement procedures are explained by the characteristics of these procedures, but measurement bias is neglected. Andrews' study was followed by several others: Rodgers, Andrews, and Herzog (1992) extended Andrews' work in USA; Költringer (1993) carried out a similar study in Austria; Scherpenzeel and Saris (1993) went further to bring together samples from several countries, although the scope of the questions was fairly limited; Alwin and Krosnick (1991) carried out a somewhat related study in which only reliability was evaluated and which used quasi-simplex models (Heise, 1969; Wiley and Wiley, 1970) rather than MTMM models. Quasi-simplex models have later been criticised (Coenders, Saris, Batista-Foguet, and Andreenkova, 1999).

In this paper we attempt to combine the study of bias, reliability and validity by integrating the evaluation of relative bias in MTMM models. With this purpose, MTMM models will be adapted into *mean-and-covariance structure models* (Sörbom, 1974) including constraints representing different forms of relative unbiasedness. First, a standard MTMM model will be reviewed. The problem of relative bias will next be introduced, together with the necessary model modifications and definitions. Finally, the procedure will be illustrated with an empirical example.

2 MTMM models

MTMM designs (Campbell and Fiske, 1959) consist of multiple measures of a set of factors (*traits*) with the same set of measurement procedures (*methods*). So,

these designs include t×m measurements, that is the number of methods (m) times the number of traits (t). The differences between methods can be any design characteristic which can be shared by measurements of all traits, such as different response scale lengths or category labels in questionnaires, different data collection procedures, different raters, etc.

Method effects can often be viewed as a form of systematic error variance which is connected to the method. So, in addition to trait variance, MTMM measurements have two sources of error variance: noise or random error variance and method variance. Since the second source of error variance is common for all measurements using the same method, the resulting error terms will be correlated.

Random measurement errors tend to attenuate the correlations among observed measurements with respect to the correlations among the trait factors. On the contrary, *correlated measurement errors* will usually increase the correlations among observed measurements in absolute value (at least if trait correlations are positive). The former are related to reliability and the latter to validity.

Campbell and Fiske (1959) suggested using MTMM designs for convergent and discriminant validation by directly examining the elements of the correlation matrix among all t×m measurements, called *MTMM matrix*. An example of such a matrix can be found in Table 1. This approach was cumbersome and often led to confusion (Schmitt and Stults, 1986) so that from the early seventies MTMM matrices began instead to be analyzed by means of *structural equation models* (see for instance Bollen, 1989 as a general reference for structural equation models and Schmitt and Stults for applications on MTMM data). These models are called MTMM models and have been used for providing the researcher with reliability and validity estimates (usually in the form of a variance decomposition into trait, error, and method variance) and corrected trait correlations, taking random and correlated measurement errors into account.

Table 1: Correlations, means and standard deviations of nine measurements of life satisfaction.

	t1-m1	t2-m1	t3-m1	t1-m2	t2-m2	t3-m2	t1-m3	t2-m3	t3-m3
t1-m1	1.000								
t2-m1	0.514	1.000							
t3-m1	0.428	0.435	1.000						
t1-m2	0.693	0.469	0.343	1.000					
t2-m2	0.464	0.764	0.380	0.568	1.000				
t3-m2	0.332	0.398	0.812	0.383	0.434	1.000			
t1-m3	0.661	0.432	0.335	0.690	0.461	0.316	1.000		
t2-m3	0.412	0.762	0.352	0.451	0.779	0.365	0.527	1.000	
t3-m3	0.297	0.351	0.802	0.310	0.344	0.823	0.365	0.373	1.000
Means	74.997	80.724	65.849	75.496	81.382	68.188	73.884	80.432	66.295
Stdev	23.464	24.759	30.107	21.743	23.082	27.403	21.438	22.157	28.529

Many different MTMM models have been suggested in the literature. Among them are the *confirmatory factor analysis* (CF) model for MTMM data (Althauser, Heberlein, and Scott, 1971; Alwin, 1974; Werts and Linn, 1970), the *correlated uniqueness* model (Kenny, 1976; Marsh, 1989; Marsh and Bailey, 1991), and the *true score* model for MTMM data (Saris and Andrews, 1991). This paper will build on the CF model, which will next be presented, although the other models may also be used if so desired.

The CF model belongs to the family of factor analysis models and is probably the model most frequently used to analyze MTMM data. In this model, each observed variable is allowed to load on both one trait factor and one method factor. The latter type of factors account for error covariances or method effects. The model is specified as follows:

$$x_{ij} = \lambda_{Tij}\,\xi_{Ti} + \lambda_{Mij}\,\xi_{Mj} + \delta_{ij} \qquad \forall\, i,j \qquad (1)$$

where x_{ij} is the measurement of Trait i with Method j, expressed in deviations from the mean, δ_{ij} is the random measurement error for x_{ij} , assumed to have a zero mean, and with variance θ_{ij}; ξ_{Ti} are the trait factors expressed in deviations from the mean, with covariances $\phi_{Tii'}$ and variances ϕ_{Tii} ; ξ_{Mj} are the method factors, expressed in deviations from the mean, with variances ϕ_{Mjj} ; and λ_{Tij} is the loading of x_{ij} on ξ_{Ti} and λ_{Mij} is the loading of x_{ij} on ξ_{Mj}.

The model is usually specified with the standard assumptions of factor analysis models, including uncorrelatedness of error terms, plus the additional one of uncorrelated method and trait factors:

$$\begin{aligned} &cov(\delta_{ij}\,\xi_{Ti'}) = 0 && \forall\, ij,i' \\ &cov(\delta_{ij}\,\xi_{Mj'}) = 0 && \forall\, ij,j' \\ &cov(\delta_{ij}\,\delta_{i'j'}) = 0 && if\ i \neq i'\ or\ j \neq j' \\ &cov(\xi_{Ti}\,\xi_{Mj}) = 0 && \forall\, i,j \end{aligned} \qquad (2)$$

where i, i',... identify the traits and j, j',... identify the methods. Note that, in all equations in the article, i may be equal to i' and j may be equal to j' unless the opposite is expressly stated.

The assumption of absence of correlation between trait and method factors makes it possible to decompose the variance of x_{ij} into variance explained by the trait ($\lambda^2_{Tij}\,\phi_{Tii}$), by the method ($\lambda^2_{Mij}\phi_{Mjj}$), and random error variance (θ_{ij}) in order to assess measurement quality (e.g., Schmitt and Stults, 1986), although this assumption may not be reasonable under certain conditions (Kumar and Dillon, 1992). The percentage of variance explained by the trait can be interpreted as the product of reliability and validity if the model is correctly specified. This product is referred to as quality of the item in Saris (1996). Reliability can be computed as one minus the percentage of random error variance.

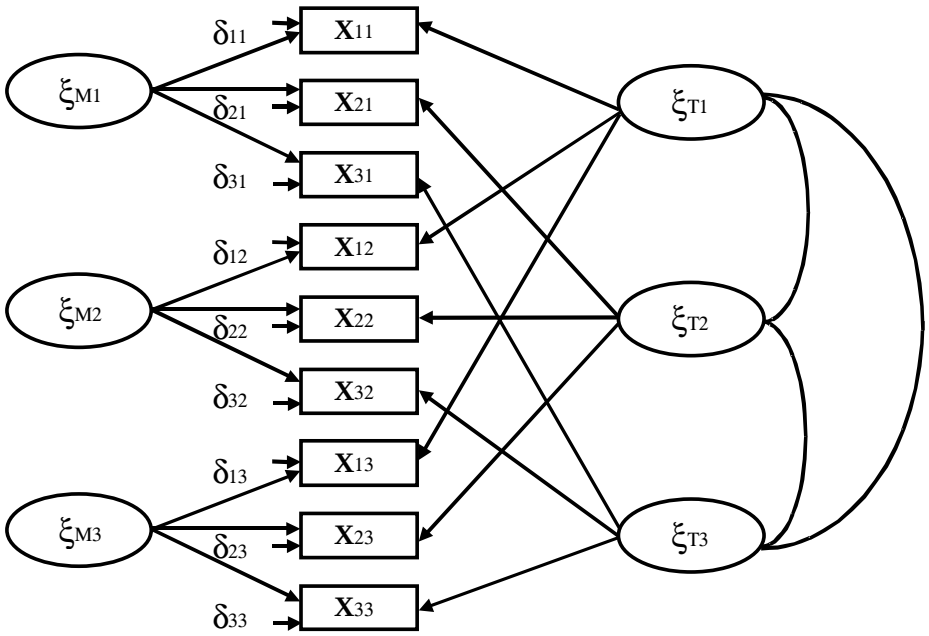

Figure 1: Path diagram of the CF model for three traits and three methods.

Two additional sets of constraints are considered in this paper in order to avoid the frequent overparametrization problems of the model, such as failure to converge, inadmissible estimates, or empirical underidentification (Andrews, 1984; Bagozzi and Yi, 1991; Brannick and Spector, 1990; Kenny and Kashy, 1992; Marsh and Bailey, 1991; Saris, 1990a).

$$\lambda_{Mij} = 1 \qquad \forall\, i, j \qquad (3)$$

$$cov(\xi_{Mj}\ \xi_{Mj'}) = 0 \qquad if\ j \neq j' \qquad (4)$$

The restriction in Equation 4 of uncorrelated method factors implies that only error covariances among indicators sharing the same method can be explained by the model. In some circumstances, some of the methods of measurement are similar, thus suggesting the existence of error covariances among measurements using different methods (de Wit, 1994). For instance, Andrews and Withey (1976, chap. 6) consider six methods for evaluating perceived satisfaction, five based on self-ratings and one on other's ratings. The authors expected correlated measurement errors to occur among all self-rating measures. In such a case the constraint in Equation 4 would not hold. Fortunately, some literature suggests that the effect of such misspecification is fairly minor (Marsh and Bailey, 1991; Saris, 1990b; Scherpenzeel, 1995), at least if method variance is low. In any case, eliminating the assumption of method uncorrelatedness leads to many problems, as mentioned above. The CF model specified with the restrictions in Equations 2 to 4

was used for instance in Andrews (1984) and Saris (1990a). Figure 1 shows a path diagram of the model for t=3 and m=3.

3 Bias assessment in MTMM designs

This paper is concerned with the assessment of bias caused by the methods. As has been said, MTMM designs and models have traditionally been used to assess reliability and validity of measurement instruments. Relative bias of measurement instruments has mainly been evaluated by using split-ballot experiments (e.g., Schuman and Presser, 1981) rather than by using structural equation models such as MTMM models.

In this article it is shown how to integrate the assessment of bias into MTMM models in order to simultaneously evaluate all three aspects of measurement quality. The correlations or covariances among the variables are in principle enough to assess reliability and validity by means of MTMM models. In order to study reliability, validity and relative bias at one go, means must also be included in the analysis and a mean-and-covariance structure model (e.g., Bollen, 1989; Jöreskog and Sörbom, 1989; Sörbom, 1974) is called for. There is another major difference between validity or reliability assessment and relative bias assessment:

1. Validity and reliability of a method can be estimated in absolute terms. In other words, their estimates for a given method should not in principle change depending on which other methods are combined with it in the MTMM design. This occurs because reliability and validity concern only the strength of the relationship between the trait factor and the observed measurement and the correlations among measurements provide enough information to identify and estimate this strength of relationship. Of course, this only holds if the MTMM model is correctly specified. Költringer (1995a), and de Wit (1994) show that absolute estimates are not possible otherwise.
2. Bias of a method can only be estimated in comparative terms with respect to another particular method. This occurs because bias of a measurement using a given method concerns the comparison of the mean of the observed measurement and the mean of the trait factor. The mean of the trait factor is completely arbitrary and can at most be fixed according to the mean of a measurement using another method. The researcher can at most assess whether the means of measurements made with different methods are the same or not, but cannot decide which is correct, if any is correct at all. Then, only statements regarding the comparison of a method with a particular alternative method are possible. Attempts have been made to provide absolute indices of bias (Költringer, 1995b) but the value of these

indices will strongly depend on the particular set of methods which are combined the MTMM design. Unlike reliability and validity estimates, bias estimates continue to be relative even if the model is correctly specified.

The fact that bias can only be assessed in relative terms does not mean that its study is useless. The study of relative bias is still necessary and useful if the results obtained in different surveys using different methods are to be compared.

3.1 Reparametrization of MTMM models to include means

In order to study reliability, validity and relative bias at one go, covariances and means must both be included in the analysis. The CF model can easily be reparametrized as a mean-and-covariance structure model deal with variables which are not centred about their means. The main difference with respect to the classic specification will be the inclusion of the means of the trait factors and intercept terms in the equations as parameters to be estimated. The means of trait factors are obviously interesting to the researcher as they intend to express the mean values of the population for the characteristics being measured. The model is then specified as follows:

$$x_{ij} = \tau_{ij} + \lambda_{Tij}\,\xi_{Ti} + \lambda_{Mij}\,\xi_{Mj} + \delta_{ij} \qquad \forall\, i,j \qquad (5)$$

where τ_{ij} is an intercept term in the equation; x_{ij} is the uncentered measurement of Trait i with Method j. The ξ_{Ti} trait factors are uncentered with means κ_i while the ξ_{Mj} method factors are expressed in deviations from their means. The remaining terms are interpreted as in Equation 1.

In order to identify the ϕ_{Tii} parameters, one λ_{Tij} loading must be constrained to 1 for each trait. Similarly, in order to identify the κ_i parameters, one τ_{ij} must be constrained to 0 for each trait. Let us assume without loss of generality that the constraint is applied to all measurements with Method 1 so that $\lambda_{Ti1}=1$ and $\tau_{i1}=0$ $\forall$ i.

The new parametrization in Equation 5 and the constraints $\lambda_{Ti1}=1$ and $\tau_{i1}=0$ fix the scale and origin of the trait factors according to their measurements with Method 1. From the reparametrized models it can then be evaluated whether Methods 2 to m lead to scores which are systematically different from those obtained with Method 1, once random errors and correlated errors or method effects have been accounted for. This stresses the comparative nature of bias assessment. Of course, it is up to the researcher to decide which of the methods in the design will be labelled Method 1.

The model parameters which are useful to evaluate bias relative to Method 1 are the τ_{ij} intercepts and the λ_{Tij} trait loadings with $j\neq1$. Bias is understood as a

difference between the scaling of the measurements and the trait factors. Method factors are, thus, not directly involved. The method factor variances express systematic response behaviour which is constant within a method but varies across the different respondents, while the τ_{ij} intercepts express a response behaviour which is constant across all respondents but may vary within a method from trait to trait; the former correspond to what is understood as a method effect and the latter can be related to measurement bias.

3.2 Alternative definitions of relative bias

Before continuing, it must be made clear what we understand by measurement bias in the context of the reparametrized model. In the literature on survey research, bias of measurement instruments is usually considered when using the responses to the instrument with the aim of estimating a parameter of interest related to the population of respondents, usually a proportion or a mean. More precisely, it is "the type of error that affects the statistic in all implementations of a survey design; in that sense it is a constant error" (Groves, 1989: 8).

Here we concentrate on one characteristic of the survey design, namely the method. If we consider the estimation of the κ_i trait means, through the mean of the x_{ij} observed scores across all respondents with Method j, then Method j is unbiased if:

$$E(mean(x_{ij})) = \kappa_i \qquad \forall i \qquad (6)$$

where the operator *mean* denotes the average across all respondents and the operator E denotes the expectation across all possible replications of the survey. Unfortunately, absolute bias as defined in Equation 6 cannot be assessed because κ_i is not observable.

This leads us to a first definition of relative unbiasedness of Methods 1 and j which states that both methods have the same bias and thus the expectation of the mean of the responses across all respondents must be the same for both methods regardless of the trait. This will be referred to as *unconditional definition* of relative unbiasedness of Methods 1 and j:

$$E(mean(x_{i1})) = E(mean\ (x_{ij})) \qquad \forall i \qquad (7)$$

It must be noted that the expectations of the means of x_{i1} and x_{ij} implied by the models are:

$$E(mean(x_{i1})) = \kappa_i \qquad \forall i \qquad (8)$$

$$E(mean(x_{ij})) = \tau_{ij} + \lambda_{Tij}\,\kappa_i \qquad \forall i \qquad (9)$$

where it can, of course, be seen that the implied mean of measurements with Method 1 equals the trait mean. The unconditional definition then implies that the model parameters must fulfil the non-linear constraint

$$\tau_{ij} + \lambda_{Tij}\,\kappa_i = \kappa_i \qquad \forall\, i \qquad (10)$$

If $\tau_{ij}=0$ and $\lambda_{Tij}=1\ \forall$ i then Methods 1 and j are relatively unbiased with respect to the above definition, but this is not the whole story. Both the loadings and the intercepts will usually have to be jointly interpreted in order to evaluate bias. Note that (assuming, without loss of generality, that κ_i is positive), a positive τ_{ij} may not imply that $E(\text{mean}(x_{ij})) > E(\text{mean}(x_{i1}))$ when $\lambda_{Tij}<1$. Similarly, a negative τ_{ij} may not imply that $E(\text{mean}(x_{ij})) < E(\text{mean}(x_{i1}))$ when $\lambda_{Tij}>1$. Only when $\lambda_{Tij}=1$ does τ_{ij} correspond to the change in the mean. The unconditional definition is thus complicated to relate to the model parameters.

Furthermore, even if $E(\text{mean}(x_{ij})) = E(\text{mean}(x_{i1}))$, the fact that λ_{Tij} deviates from 1 may have a substantive interest. If $\lambda_{Tij}>1$ and $\tau_{ij}<0$, then the respondents tend to give more extreme or polarized answers with Method j than with Method 1. If $\lambda_{Tij}<1$ and $\tau_{ij}>0$, then the respondents tend to give answers which are closer to the mean with Method j than with Method 1. The λ_{Tij} parameter relates the standard deviation of x_{ij} to the standard deviation of x_{i1}, once method and error variance have been subtracted. Since the standard deviation of the raw responses is affected by the error and method variances, this type of conclusions cannot be drawn from split-ballot experiments.

In order to simplify the relationship between the definition and the model parameters and in order to take standard deviations corrected for measurement error into account, we suggest using a more strict alternative definition of relative unbiasedness in which the parameters of interest to be estimated are the scores of each respondent on the trait in a psychometric sense (see Groves, 1989, Cap. 1 for a discussion about the different perspectives of bias across disciplines). Methods 1 and j are relatively unbiased if the expectation of the responses conditional on the value of ξ_{Ti} is the same for both methods regardless of the trait:

$$E(x_{i1}/\xi_{Ti}=k) = E(x_{ij}/\xi_{Ti}=k) \qquad \forall\, i,k \qquad (11)$$

This definition will be referred to as conditional definition. Assuming independence between trait and method factors, these conditional expectations can be expressed as:

$$E(x_{i1}/\xi_{Ti}=k) = k \qquad \forall\, i,k \qquad (12)$$

$$E(x_{ij}/\xi_{Ti}=k) = \tau_{ij} + \lambda_{Tij}\ k \qquad \forall\, i,k \qquad (13)$$

In order that Methods 1 and j be relative unbiased, the condition below must hold *for all possible values* of k.

$$k = \tau_{ij} + \lambda_{Tij} \quad k \qquad \forall\, i,k \qquad (14)$$

The latter will imply that $\tau_{ij}=0$ and $\lambda_{Tij}=1$. It must be noted that relative unbiasedness according to the unconditional definition is a necessary but not sufficient condition for relative unbiasedness according to the conditional definition. The conditional definition will be the one considered in this article.

3.3 Estimating and testing mean-and-covariance MTMM models

The estimation of mean-and-covariance structure MTMM model in Equation 5 can be carried out with most standard software for structural equation modeling if an augmented moment matrix or a covariance matrix and a mean vector are supplied. The estimation is usually be made by normal-theory maximum likelihood. Satorra (1992) showed that some maximum likelihood inferences are asymptotically robust to non-normality as long as the random constituents of the model (δ_{ij} error terms, ξ_{Ti} trait factors and ξ_{Mj} method factors) which are nonnormal fulfil two conditions, namely they have unconstrained variances and are either mutually independent or have unconstrained covariances. If these conditions hold, the likelihood ratio χ^2 tests are correct. Moreover, the standard errors of the estimates is also correct except for variance and covariance parameters of non normal random constituents of the model. The models used in this article do not introduce any constraints on variances and the tests considered involve loadings and intercepts only. We thus expect inferences to be robust.

3.4 Assessing consistency of the behaviour of methods

A set of nested models can be tested in order to evaluate relative bias from the pattern and the statistical significance of the τ_{ij} and λ_{Tij} parameters for j=2,...,m.

The suggested strategy starts by assessing whether these parameters are constant within a method over all traits. If this holds, then a constant behaviour of the methods in terms of relative bias is supported. If this does not hold, then the relative bias of a method may change from trait to trait, and no general conclusions about the behaviour of the method can be drawn: an interaction between trait and method is instead supported. The following constraints must then be tested:

$$\tau_{ij} = \tau_{i'j} = \tau_j \qquad \forall\, i \neq i',\ j=2,...,m \qquad (15)$$

$$\lambda_{Tij} = \lambda_{Ti'j} = \lambda_{Tj} \qquad \forall\, i \neq i',\ j=2,...,m \qquad (16)$$

Three constrained models can be considered in which the constraints in Equations 15 and 16 are applied isolatedly or jointly.

3.5 Assessing relative bias of methods

If the constraints in the foregoing subsection are not rejected, then the behaviour of the methods can be considered to be approximately stable and it can next be assessed whether this stable behaviour is significantly biased with respect to Method 1. This can be done by testing the constraints:

$$\tau_j = 0 \qquad j = 2,...,m \qquad (17)$$

$$\lambda_{Tj} = 1 \qquad j = 2,...,m \qquad (18)$$

These constraints can be introduced for a single method, any subset of methods or all methods simultaneously.

4 Illustration

In this section, the mean-and-covariance structure models which have been presented will be fitted to real data from a survey of life satisfaction. The constraints which have been presented in previous sections will then be orderly tested in order to assess the presence of relative bias.

4.1 Data

The data from a survey of perceived life satisfaction carried out in the greater Göteborg area (Sweden) in Autumn 1989 will be used for illustration. We consider t=3 domains of life satisfaction (traits):

1. Life in general (t1).
2. Housing situation (t2).
3. Financial situation (t3).

Each trait was measured with m=3 response scales ranging from “completely dissatisfied” to “completely satisfied” (methods):

1. 1 to 100 line production scale in which respondents were asked to draw a line whose length was then measured from 1 to 100 (m1).
2. 1 to 10 numeric scale (m2).

3. 1 to 5 scale with all-labelled categories (m3).

The questionnaire was administered by mail in a three-wave panel design. In each wave one method was presented. The time interval between the waves was about two weeks and the order in which the methods were administered varied for different groups of respondents. The sample size was N=336 after applying listwise deletion of missing values. Further details on the questionnaire and data collection can be found in Olsen and Munck (1996). The wording of the items and the order in which the traits were presented was the same for all methods: t1, t2, t3. However, the three methods themselves are sufficiently different (line production versus number and categorical) so that some relative bias may be expected.

4.2 Converting measurements to a common scale

For this data set, the different methods produce measurements expressed in different units. Prior to assessing relative bias, the measurements of the same trait obtained using different methods must be converted to a common or comparable scale.

One possible transformation is to rescale all measurements so that they have a common allowed range, for instance by letting the minimum *allowed* response be 0 and the maximum allowed response be 100:

$$\textit{rescaled response} = \frac{\textit{response - allowed minimum}}{\textit{allowed maximum - allowed minimum}} \times 100 \qquad (19)$$

The changes of scale do not imply that an equivalence be necessarily imposed between the units of a measurement scale and another. The λ_T parameters can still account for differences in this respect if necessary: precisely these parameters will be a powerful instrument for the assessment of relative bias.

Non-linear rescaling transformations could have been considered as an alternative. We disregarded them because they not only affect the means and variances but also the correlations.

The Pearson correlation matrix and the means and standard deviations of the nine measurements in our data set are given in Table 1 and refer to the rescaled variables to a 0 to 100 range.

Table 2: Estimates of the CF model. Unconstrained and with constraints "I" in Table 3.($\tau_{ij}=\tau_j$ and $\lambda_{ij}=\lambda_j$; i=1,2,3; j=2,3; $\tau_3=0$; $\lambda_3=1$).

Measurement equations

	Unconstrained		Constrained	
	τ	λ_T	τ	λ_T
t1-m1	0.00	1.00	0.00	1.00
t2-m1	0.00	1.00	0.00	1.00
t3-m1	0.00	1.00	0.00	1.00
t1-m2	2.68	0.97	3.84	0.97
t2-m2	4.16	0.96	3.84	0.97
t3-m2	6.49	0.94	3.84	0.97
t1-m3	2.98	0.95	0.00	1.00
t2-m3	5.45	0.93	0.00	1.00
t3-m3	2.36	0.97	0.00	1.00

Variances, covariances and means of trait factors

	t1	t2	t3	t1	t2	
t1	357.16			343.46		
t2	250.44	451.37		236.85	422.32	
t3	218.16	258.28	709.13	208.27	242.84	678.17
means	75.00	80.72	65.85	74.30	80.46	66.29

Method variances

	m1	m2	m3	m1	m2	m3
	50.63	33.23	40.02	52.89	34.88	38.82

Error variances

	m1	m2	m3	m1	m2	m3
t1	131.64	96.96	106.60	134.05	100.59	101.51
t2	112.85	79.96	70.55	116.21	83.66	64.53
t3	135.57	96.38	119.38	139.56	93.83	118.00

Goodness of fit

	χ^2	d.f.	χ^2	d.f.
	17.22	21	24.45	31

4.3 Assessing consistency of the behaviour of methods

The first column of Table 2 presents the maximum likelihood estimates and goodness of fit test statistic for the unconstrained mean-and-covariance structure CF model fitted to the data in Table 1. The LISREL8 program (Jöreskog and Sörbom, 1989, 1993) was used. The model yields a χ^2 statistic of 17.22 with 21 degrees of freedom and is thus not rejected, which allows us to proceed to test the constraints of relative unbiasedness.

As has been argued, in order to determine whether a method produces some systematic kind of bias it must first be assessed whether the behaviour of the method is any systematic or stable at all. The tests of stability of method behaviour are illustrated on the data in Table 1. The first row of Table 3 shows the χ^2 statistics of the model constrained to a stable behaviour of methods. From Table 3 and Table 2, the corresponding χ^2 change statistics needed to perform the tests can be computed. From these tests, the constraint $\tau_{ij}=\tau_j$; i=1,2,3; j=2,3 cannot be rejected (the χ^2 change between the unconstrained model and the model labelled "A" in Table 3 is 2.32; with 4 degrees of freedom); the constraint $\lambda_{Tij}=\lambda_{Tj}$; i=1,2,3; j=2,3 cannot be rejected (the χ^2 change between the unconstrained model and the model labelled "B" in Table 3 is 1.78; with 4 degrees of freedom); and both sets of restrictions cannot either be jointly rejected (the χ^2 change between the unconstrained model and the model labelled "C" in Table 4 is 4.08; with 8 degrees of freedom). This result supports the consistency of the behaviour of all methods.

4.4 Assessing relative bias of methods

On the basis of model "C" in Table 3, the tests of relative bias of m2 with respect to m1 can be performed by comparing its fit to that of the models "D" to "F". Although none of the models is statistically rejected based on the overall χ^2 test, from the corresponding χ^2 change tests, the assumption $\tau_2=0$ can be rejected at α=1% (χ^2 change=6.58, with 1 d.f.), the assumption $\lambda_{T2}=1$ can be rejected at α=3% (χ^2 change=4.70, with 1 d.f.) and both assumptions can also be jointly rejected at α=2.5% (χ^2 change=7.40, with 2 d.f.). According to these results, the relative unbiasedness of m1 and m2 is not tenable according to the conditional definition.

On the basis of model "C" in Table 3, the tests of relative bias of m3 with respect to m1 can be performed by comparing its fit to that of the models "G" to "I". From the corresponding χ^2 change tests, the assumption $\tau_3=0$ cannot be rejected (χ^2 change=2.22, with 1 d.f.), the assumption $\lambda_{T3}=1$ cannot either be rejected under the usual risk standards (χ^2 change=2.94, with 1 d.f., α=8.6%) and both assumptions cannot either be jointly rejected (χ^2 change=3.15, with 2 d.f.).

According to these results, the relative unbiasedness of m1 and m3 is tenable, both conditionally and unconditionally.

Table 3: Goodness of fit of constrained mean-and-covariance structure CF models.

Testing stability of method behaviour

Label	A	B	C
Constraints	$\tau_{ij}=\tau_j$;i=1,2,3;j=2,3	$\lambda_{Tij}=\lambda_{Tj}$;i=1,2,3;j=2,3	$\tau_{ij}=\tau_j$;i=1,2,3;j=2,3 $\lambda_{Tij}=\lambda_{Tj}$;i=1,2,3;j=2,3
χ^2 statistic	19.54	19.00	21.30
Degrees of freedom	25	25	29
p value	0.7705	0.7970	0.8480

Testing relative bias of Methods 2 and 1

Label	D	E	F
Constraints	$\tau_{ij}=\tau_j$;i=1,2,3;j=2,3 $\lambda_{Tij}=\lambda_{Tj}$;i=1,2,3;j=2,3 $\tau_2=0$	$\tau_{ij}=\tau_j$;i=1,2,3;j=2,3 $\lambda_{Tij}=\lambda_{Tj}$;i=1,2,3;j=2,3 $\lambda_{T2}=1$	$\tau_{ij}=\tau_j$;i=1,2,3;j=2,3 $\lambda_{Tij}=\lambda_{Tj}$;i=1,2,3;j=2,3 $\tau_2=0$, $\lambda_{T2}=1$
χ^2 statistic	27.88	26.00	28.70
Degrees of freedom	30	30	31
p value	0.5769	0.6753	0.5849

Testing relative bias of Methods 3 and 1

Label	G	H	I
Constraints	$\tau_{ij}=\tau_j$;i=1,2,3;j=2,3 $\lambda_{Tij}=\lambda_{Tj}$;i=1,2,3;j=2,3 $\tau_3=0$	$\tau_{ij}=\tau_j$;i=1,2,3;j=2,3 $\lambda_{Tij}=\lambda_{Tj}$;i=1,2,3;j=2,3 $\lambda_{T3}=1$	$\tau_{ij}=\tau_j$;i=1,2,3;j=2,3 $\lambda_{Tij}=\lambda_{Tj}$;i=1,2,3;j=2,3 $\tau_3=0$, $\lambda_{T3}=1$
χ^2 statistic	23.52	24.24	24.45
Degrees of freedom	30	30	31
p value	0.7932	0.7610	0.7918

To sum up, the stability of behaviour of methods across traits, and the unbiasedness of m3 (1 to 5 scale) with respect to m1 (1 to 100 line production scale) are maintained while the unbiasedness of m2 (1 to 10 scale) with respect to m1 is rejected.

4.5 Assessing overall measurement quality

The model with restrictions "I" in Table 3 is then considered for interpretation. This model constrains all trait loadings for m1 and m3 to 1 and all intercepts for m1 and m3 to 0. It also constrains trait loadings and intercepts with m2 to be equal across measurements of all traits. Note the high degree of parsimony of the model (31 d.f., compared to the 21 d.f. achieved by the original model) which has been achieved by introducing substantively meaningful constraints, whose failure to be rejected also provided interesting conclusions with respect to measurement quality. Parsimony leads both to ease of interpretation and efficiency: the estimated standard errors of the estimates dropped for nearly all parameters, in some cases by more than half.

The actual estimates of all parameters of the model with the constraints "I" are in the second column of Table 2. Some of these estimates (trait variances, covariances and means) refer to the distribution of the trait factors corrected for random measurement errors and method effects and are expressed in the units of the line production scale.

The remaining estimates allow us to draw a complete picture of measurement quality and include parameters related to the relative bias of the line production (m1) and numeric (m2) scales, (loadings and intercepts of m2); and parameters related to reliability and validity (method variances and error variances). These estimates show that the ten-point scale (m2) tends to yield measurements which are less polarized (i. e. closer to the mean) than the remaining methods. They also allow us to compute the percentages of trait, random error and method variance for each of the 9 measurements. According to Section 2, reliability can then computed from the percentage of random error variance, and validity from reliability and the percentage of trait variance. These measurement quality estimates are displayed in Table 4. The line production scale (m1) yields the lowest reliability and validity. Reliability and validity are also lowest for measurements of satisfaction with life in general (t1).

5 Discussion

This article was concerned with the assessment of bias caused by the methods of measurement. As has been said, MTMM designs and models have traditionally been used to assess reliability and validity of measurement instruments. Relative bias of measurement instruments has mainly been evaluated by using split-ballot experiments (e.g., Schuman and Presser, 1981) rather than by using structural equation models such as MTMM models. This has resulted in:

Table 4: Measurement quality from the CF model with constraints "I" in Table 3.

($\tau_{ij}=\tau_j$ and $\lambda_{ij}=\lambda_j$; i=1,2,3; j=2,3; $\tau_3=0$; $\lambda_3=1$)

	Percentages of Measurement variance explained by:			quality	
	trait	random error	method	reliability	validity
t1-m1	64.8%	25.3%	10.0%	74.7%	86.7%
t2-m1	71.4%	19.6%	8.9%	80.4%	88.9%
t3-m1	77.9%	16.0%	6.1%	84.0%	92.8%
t1-m2	70.5%	21.9%	7.6%	78.1%	90.3%
t2-m2	77.0%	16.2%	6.8%	83.8%	91.9%
t3-m2	83.2%	12.2%	4.5%	87.8%	94.8%
t1-m3	71.0%	21.0%	8.0%	79.0%	89.8%
t2-m3	80.3%	12.3%	7.4%	87.7%	91.6%
t3-m3	81.2%	14.1%	4.6%	85.9%	94.6%

1. Only some aspects of measurement quality are evaluated at a time, while there may be interrelations between them and all of them are necessary to compare the results of different surveys.
2. Only means are used to evaluate relative bias, which implies a quite restrictive definition of bias.

In this article it was shown how to integrate the assessment of bias into MTMM models in order to simultaneously evaluate all three aspects of measurement quality. While bias continues to be treated in a relative way under this approach, as the case was in split-ballot experiments, the definition of relative bias has been enriched and complemented with statistical tests based on model constraints.

An illustration of the tests has been provided in which relative unbiasedness was tenable for certain pairs of methods and rejected for certain others. In spite of the generality of the approach, it can lead to the fit of very parsimonious models, as shown in our example, in which a multitrait-multimethod model analyzing a design with only 9 variables left 31 degrees of freedom.

References

[1] Althauser, R.P., Heberlein, T.A., and Scott, R.A. (1971): A causal assessment of validity: The augmented multitrait-multimethod matrix. In H. M. Blalock, Jr. (Ed.), *Causal Models in the Social Sciences.* Chicago: Aldine, 151-169.

[2] Alwin, D. (1974): An analytic comparison of four approaches to the interpretation of relationships in the multitrait-multimethod matrix. In H.L.

Costner (Ed.), *Sociological Methodology 1973-1974*. San Francisco: Jossey-Bass, 79-105.

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

[4] Andrews, F.M. (1984): Construct validity and error components of survey measures. A structural modeling approach. *Public Opinion Quarterly*, **48**, 409-442.

[5] Andrews, F.M. and Withey, S.B. (1976): *Social Indicators of Well-Being. Americans' Perceptions of Life Quality*. New York: Plenum Press.

[6] Bagozzi, R.P. and Yi, Y. (1991): Multitrait-multimethod matrices in consumer research. *Journal of Consumer Research*, **17**, 426-439.

[7] Billiet, J., Loosveldt, G., and Waterplas, L. (1986): *Het Survey-Interview Onderzocht: Effecten van het Ontwerp en Gebruik van Vragenlijsten op de Kwaliteit van de Antwoorden* [Research on surveys: effects of the design and use of questionnaires on the quality of the responses] Leuven: Sociologisch Onderzoeksinstituut KU Leuven.

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

[9] Brannick, M.T. and Spector, P.E. (1990): Estimation problems in the block-diagonal model of the multitrait-multimethod matrix. *Applied Psychological Measurement*, **14**, 325-339.

[10] Campbell, D.T. and Fiske, D.W. (1959): Convergent and discriminant validation by the multitrait multimethod matrices. *Psychological Bulletin*, **56**, 81-105.

[11] Coenders, G., Saris, W.E., Batista-Foguet, J.M., and Andreenkova, A. (1999): Stability of three-wave simplex estimates of reliability. *Structural Equation Modeling*, **6**, 135-157.

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

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

[14] Groves, R.M. (1989): *Survey Errors and Survey Costs*. New York: John Wiley & Sons.

[15] Jöreskog, K.G. and Sörbom, D. (1989): *LISREL7, A Guide to the Program and Applications*. Chicago: SPSS Inc.

[16] Jöreskog, K.G. and Sörbom, D. (1993): *New Features in LISREL8*. Chicago: Scientific Software International.

[17] Kenny, D.A. (1976): An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology*, **12**, 247-252.

[18] Kenny, D.A. and Kashy, D.A. (1992): Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, **112**, 165-172.

[19] Költringer, R. (1993): *Messqualität in der Sozialwissenschaftlichen Umfrageforshung* [Measurement quality in survey research in the social sciences]. Endbericht Project P8690-SOZ. Viena: Fonds zur Förderung der Wissenschaftlichen Forshung (FWF).

[20] Költringer, R. (1995a): Categorization and measurement quality. A population study using artificial multitrait-multimethod data. In W.E. Saris and Á. Münnich (Eds.), *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments*. Budapest, Hungary: Eötvös University Press, 103-124.

[21] Költringer, R. (1995b): Measurement quality in Austrian personal interview surveys. In W.E. Saris and Á. Münnich (Eds.), *The Multitrait-Multimethod Approach to Evaluate Measurement Instruments*. Budapest, Hungary: Eötvös University Press, 207-224.

[22] Kumar, A. and Dillon, W.R. (1992): An integrative look at the use of additive and multiplicative covariance structure models in the analysis of MTMM data. *Journal of Marketing Research*, **29**, 51-64.

[23] Marsh, H.W. (1989): Confirmatory factor analysis of multitrait-multimethod data: many problems and few solutions. *Applied Psychological Measurement*, **13**, 335-361.

[24] Marsh, H.W. and Bailey, M. (1991): Confirmatory factor analyses of multitrait-multimethod data: comparison of the behavior of alternative models. *Applied Psychological Measurement*, **15**, 47-70.

[25] Molenaar, N.J. (1986): *Formulieringseffecten in Survey-Interviews* [formulation effects in survey interviews]. Amsterdam: Vrije Universiteit Uitgeverij.

[26] Olsen, A. and Munck, I.M.E. (1996): Satisfaction in two Swedish towns. In W.E. Saris, R. Veenhoven, A. C. Scherpenzeel and B. Bunting (Eds.), *A Comparative Study of Satisfaction with life in Europe*. Budapest, Hungary: Eötvös University Press, 145-154.

[27] Rodgers, W.L., Andrews, F.M., and Herzog, A.R. (1992): Quality of survey measures: a structural modeling approach. *Journal of Official Statistics*, **8**, 251-275.

[28] Saris, W.E. (1990a): The choice of a model for evaluation of measurement instruments. In W. E. Saris and A. van Meurs (Eds.), *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait Multimethod Matrices*. Amsterdam: North Holland, 118-129.

[29] Saris, W.E. (1990b): 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 Matrices*. Amsterdam: North Holland, 52-80.

[30] Saris, W.E. (1996): Ten years of interviewing without interviewers: the telepanel. *Paper presented at the INTER-CASIC conference*, San Antonio, TX, December 1996.

[31] Saris, W.E. and Andrews, F.M. (1991): Evaluation of measurement instruments using a structural modeling approach. In P. P. Biemer, R.M. Groves, L.E. Lyberg, N. Mathiowetz, and S. Sudman (Eds.), *Measurement Errors in Surveys*. New York: John Wiley & Sons, 575-599.

[32] Satorra, A. (1992): Asymptotic robust inferences in the analysis of mean and covariance structures. In P.V. Marsden (Ed.), *Sociological Methodology 1992*. Oxford: Basil Blackwell, 249-278.

[33] Scherpenzeel, A.C. (1995): *A Question of Quality. Evaluating Survey Questions by Multitrait-Multimethod Studies*. Doctoral dissertation, University of Amsterdam. Leidschendam, the Netherlands: Royal PTT Nederland.

[34] Scherpenzeel, A.C. and Saris, W.E. (1993): The quality of indicators of satisfaction across Europe. A meta-analysis of multitrait-multimethod studies. *Bulletin de Methodologie Sociologique*, **39**, 3-19.

[35] Schmitt, N. and Stults, D.N. (1986): Methodology review. Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, **10**, 1-22.

[36] Schuman, H. and Presser, S. (1981): *Questions and Answers in Attitude Surveys: Experiments on Question Form, Order and Context*. New York: Academic Press.

[37] Sörbom, D. (1974): A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology*, **27**, 229-239.

[38] Sudman, S. and Bradburn, N.M. (1982): *Asking Questions: A practical Guide to Questionnaire Design*. San Francisco: Jossey Bass.

[39] Werts, C.E. and Linn, R.L. (1970): Path analysis. Psychological examples. *Psychological Bulletin*, **74**, 193-212.

[40] Wiley, D.E. and Wiley, J.A. (1970): Estimating measurement error using multiple indicators and several points in time. *American Sociological Review*, **35**, 112-117.

[41] de Wit, H. (1994): *Cijfers en hun achterliggende realiteit. De MTMM-kwaliteitsparameters op hun kwaliteit onderzocht* [Numbers and their underlying reality. The MTMM quality parameters for measurement quality research]. Doctoral dissertation, Catholic University of Leuven, Belgium.