KM 39 (1992)
pg.19-38

Effects of response scale, order and position on data quality: An illustration of the evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies

Annette C. Scherpenzeel & Willem E. Saris*

Abstract

In an international comparative research project, in which 15 countries cooperate, a multitrait-multimethod approach is used to get an evaluation of survey measurement instruments. In this paper, the procedure that is followed in this project is clarified and illustrated with a study of the effects of response scale, order of the response scales and position of the questions in the questionnaire, on the quality of the obtained data.

This paper was presented at the XII World Congress of Sociology. Madrid, July 9 - 13 1990.

* University of Amsterdam, Methodology Department, Faculty of Political Sciences, Oudezijds Achterburgwal 237, 1012 DL Amsterdam, The Netherlands.

1. Introduction

There is an increasing interest in the quality of measurement instruments for survey research. The quality of survey data can be affected by many characteristics of these instruments, like the length of the question and the introduction to the question, the form of the response scale, the number and labelling of response categories, the position and context of a question and the data collection technique. An approach that is often used to evaluate measurement instruments is the *multitrait-multimethod* design. This approach was first introduced by Campbell and Fiske (1959), who suggested to measure each of a number of traits with a number of different methods, and provided guidelines to infer convergent and discriminant validity directly from the multitrait-multimethod correlation matrix.

In recent work, confirmatory factor analysis is used mostly to analyse multitrait-multimethod data and to estimate validity, reliability, and method effects. Andrews (1984) first used this approach in a large scale study to evaluate many different measurement instruments across American and Canadian surveys. Very recently, a modified version of the causal model used to analyse multitrait-multimethod matrices with confirmatory factor analysis has been developed (Saris and Andrews, 1990). This second order factor model is used in an international comparative research project that is carried out at the moment. In this international project, 15 countries are collecting multitrait-multimethod data, and a meta-analysis will be carried out across all these datasets, to get an evaluation of measurement instruments in the line of Andrews' study. The study reported in this paper is part of this large scale project, and is an illustration of the procedure that will be followed. It examines the effects of the factors: response scale, order of the scales and position in the questionnaire, on the quality of the obtained data.

2. The multitrait-multimethod model

When a multitrait-multimethod (MTMM) design is used to evaluate the quality of measurement instruments, a causal model is usually specified for the vector of observed variables x as a function of t trait and m method factors, and confirmatory factor analysis is performed to analyse the data, for example with the LISREL program (Joreskog & Sorbom, 1983). Within the LISREL- framework, the causal model can be formulated as follows:

$$x = \Lambda\xi + \delta \qquad (1)$$

where: x is a vector of observed variables
$\Lambda = (\Lambda^t \quad \Lambda^m)$ is a matrix of factor loadings (mt * t+m)
$\xi' = (\xi^t \quad \xi^m)$ is a vector of t trait and m method factors (t+m * 1)
δ is a vector of residuals (mt * 1)

It is assumed that both the observed and latent variables are measured in deviations from their means and that the residuals are uncorrelated with each other and the factors:

$$E(x) = 0;\ E(\xi) = 0;\ E(\delta) = 0; \qquad (2)$$
$$E(\delta\delta') = \text{diag};\ E(\xi\delta') = 0$$

The observed covariance or correlation matrix then is:

$$E(xx') = \Sigma = \Lambda\Phi\Lambda' + \Theta_\delta \qquad (3)$$

where: $\Phi = E(\xi\xi')$ is the covariance matrix of the t + m factors
Θ_δ is the diagonal matrix of residual variances

Matrix Φ can be partitioned as:

$$\Phi = \begin{pmatrix} \Phi^{t\,t} & \Phi^{t\,m} \\ \Phi^{mt} & \Phi^{mm} \end{pmatrix} \qquad (4)$$

where: $\Phi^{t\,t}$ is a (t x t) symmetric submatrix of Φ that contains the trait factor variances and covariances or correlations
Φ^{mm} is a (m x m) symmetric submatrix of Φ that contains method factor variances and covariances or correlations
$\Phi^{mt} = \Phi^{t\,m}$ is a (m x t) rectangular submatrix of Φ that contains covariances or correlations of the m method factors with the t trait factors.

Each observed variable x_j is analyzed in this model as:

$$x_j = \lambda^t_j \xi^t + \lambda^m_j \xi^m + \delta_j \qquad (5)$$

The estimation of this confirmatory factor analysis model can sometimes lead to identification or convergence problems. As is discussed in Saris & van Meurs (1990), the most acceptable way to solve these problems is to assume zero correlations between the method factors and zero trait-method correlations (Φ^{mm} is a diagonal or identity

matrix and Φ^{mt} is a null matrix), and to assume that all method effects of a given method are equal over different traits (constraining the parameters in each column of Λ^m to be equal). These assumptions were also made by Andrews in his original study (Andrews, 1984). In the matrix of factor loadings Λ the validity coefficients and method effects can then be found. The submatrix Λ^t contains zeros except for the loadings of observed variables on their respective trait factors. The standardized factor loading λ^t_{11}, for example, of observed score x_{11} on trait factor 1, is defined as the validity coefficient of that measure: it represents the direct effect of the trait on the observed variable. The submatrix Λ^m also contains zeros, except for the loadings of the observed variables on their respective method factors. The standardized loading λ^m_{11} of the observed score x_{11} on method factor 1 is defined as the method effect coefficient of that measure: it represents the effect of that specific method on the observed variable. The validity coefficient squared is equal to the proportion valid variance, that is: the variance in the observed score, explained by the trait factor. The method effect coefficient squared is equal to the method specific variance in the observed variable. The reliability of a measure is defined in this model as 1 minus the error variance: it represents the direct effect that all variables, except the δ, have on x_j (Bollen 1989). The modified causal model that has been proposed by Saris and Andrews (Saris & Andrews, 1991) is different from the standard MTMM model by the fact that a distinction is made between true scores and observed scores. The true scores represent the stable part of the response variables corrected for random measurement error. This model, which will be called the "true score model", is a second-order factor analytic model, which can be formulated as follows:

$$y = \Lambda\eta + \varepsilon \qquad (6)$$

and:

$$\eta = \Gamma\xi + \zeta \qquad (7)$$

where: y is the vector of observed variables

η is the vector of true scores: the stable part of the observed variables, corrected for random measurement error

Λ here is a diagonal matrix with the unstandardized loadings of the observed scores on the true scores (mt * mt)

ε is a vector of random measurement errors (mt * 1)

$\xi' = (\xi^t \quad \xi^m)$ is defined as in (1)

$\Gamma = (\Gamma^t \quad \Gamma^m)$ is a matrix with the second-order factor loadings of the true scores on the t trait and m method factors (mt * m+t)

ζ is a vector of unique components for the η, the true scores

The following standard assumptions are made:

$$E(y) = 0;\ E(\eta) = 0;\ E(\varepsilon) = 0;\ E(\xi) = 0;\ E(\zeta) = 0; \qquad (8)$$
$$E(\eta\varepsilon') = 0;\ E(\eta\zeta') = 0;\ E(\xi\varepsilon') = 0;\ E(\xi\zeta') = 0;$$
$$E(\zeta\varepsilon') = 0;\ E(\varepsilon\varepsilon') = \text{diag}$$

The covariance matrix for the observed variables is:

$$E(yy') = \Sigma = \Lambda\Gamma\Phi\xi\Gamma'\Lambda' + \Lambda\Psi\Lambda' + \Theta_\varepsilon \qquad (9)$$

where: Φ is the covariance matrix of the (t + m) second- order factors and is defined as in (4)

Ψ is the covariance matrix of unique components in the true scores (mt * mt)

Θ_ε is a diagonal matrix of the random error variances (mt * mt)

For this true score model, the same additional assumptions are made as for the standard model: Φ^{mm} is a diagonal, Φ^{mt} is a null matrix and the parameters in each column of Λ^m are equal. One extra assumption is necessary to estimate the true score model: the unique components of the true scores are assumed to be zero: $E(\zeta\zeta') = 0$. Without this assumption, the model is not identified: as none of the measurement instruments is used at more than one point in time, the unique components cannot be distinguished from the random errors. The assumption is realistic only if one makes sure that different measures that are meant to assess one specific trait, do in fact measure the same construct. If, for example, one is interested in the number of response categories as the method aspect in a particular MTMM-study, then all questions used to measure the same trait in this study should be formulated literally the same and differ only in the number of response categories. If this is taken into account when designing a MTMM-study, and the questions are carefully formulated, the assumption seems reasonable. The true score validity is then defined as the standardized loading of a true score on a trait factor, and the true score method effects are defined as the standardized loadings of the true scores on the method factors. Thus in the Γ matrix both the validity coefficients and method effects can be found. The reliability coefficient of a measure is given in this case by the standardized loading of an observed score on its stable part or "true score", and can thus be found in the Λ matrix. Each observed variable y_j is analyzed in the true score model as:

$$y_j = \Lambda^t_j \Gamma^t_j \xi^t + \Lambda^m_j \Gamma^m_j \xi^m + \varepsilon_j \qquad (8)$$

Path diagrams of the standard model as it as described above and of the true score model are given in figure 1 and figure 2 in appendix A. The standard model and the true score model are mathematically equivalent: given the assumption of zero unique variance of the true scores, coefficients resulting from one of the parameterizations can always be recalculated to obtain the coefficients that would have resulted from the same data with the other parameterization. It can be seen that by multiplying the validity and reliability coefficients of the true score model, we will get the validity coefficients estimated by the standard model, and by multiplying the method effect and reliability coefficient of the true score model, we will get the method effect validity and method effect coefficients of the standard model. The reliability coefficient of the true score model squared is equal to the reliability estimate of the standard model and both parameterizations will have exactly the same goodness-of-fit. The crucial difference between the two models is the interpretation of the estimates. The standard model provides estimates of the direct effects of the latent variables on the observed variables, called "indicator validity" and the "attenuated method effect" by Saris and Andrews (1991) and the explained variance of the observed variables, which is also called the homogeniety reliability (Andrews 1984, Bollen 1989). These estimates are useful if one is interested in the (causal) relationships between certain concepts, corrected for random and systematic measurement error. The parameters of the true score model are more useful for the evaluation of measurement instruments however, because the validity coefficients and method effects as estimated by the true score model are simpler statistical quantities. The indicator validity and attenuated method effect are not independent of random measurement error. Saris and Andrews (1991) described what consequences this has for the interpretation of the validity estimates: When the true score validity is 1, we know that the true score of a measure correlates perfectly with the trait factor and can only differ from this latent factor by a scale transformation. This means that the observed variable and the trait factor are the same except for random measurement error and a possible linear scale transformation. When the true score validity is less than 1, we know exactly what the relationship between the observed variable and the latent trait factor is after correction for random measurement error. On the other hand, the indicator validity estimated by the standard model can never be 1, because there is always random measurement error. When this indicator validity is less than 1 it's unclear wether this is caused by invalidity or by unreliability (random measurement error). This shows that the true score validity is a better quantitive estimate if one is interested in specific effects on the validity of the observed variable, independent of random measurement error. The same argument holds for the interpretation of the method effect estimates. As the main purpose of the international

measurement project is to estimate and compare the specific effects of characteristics of survey instruments on validity, method effects and reliability, the true score model is more useful for this project and for the present study. For other purposes, it will always be possible to derive the standard model parameters from the estimated true score parameters.

3. The meta-analysis

In the next three years data will be collected in 15 countries, using the MTMM-design. The analyses of these datasets with the true score parameterization will result in a large number of quality estimates. The next step in the project is to carry out an analysis with these estimates as dependent variables, to explain their variation by the characteristics of the different survey questions. We call this a "meta-analysis" because it involves the analysis of quality estimates obtained from the prior causal modeling analysis. For this meta-analysis, a database has to be constructed in which one "case" consists of (codes for) all the relevant characteristics of a specific question that was used in a survey in one of the participating countries; and the estimated validity, method, and reliability coefficients of that question. Several kinds of analysis techniques can be used to analyse this database and find the effects of characteristics like response scale, position in the questionnaire or length of the questions, on the quality coefficients. Given the nature of these characteristics however, an analysis technique has to be used that can handle nominal-scale predictors, nonlinear relationships, correlations among predictors and interactive effects. Suitable techniques are therefore Multiple Classification Analysis (Andrews et al.,1973) and dummy variable multiple regression.
For the meta-analysis of the international measurement project the Multiple Classification technique, which is more convenient than the dummy variable regression approach and provides useful tables with the magnitudes of the effects of each category within a factor, was chosen. In the rest of this paper we will clear the whole procedure by an illustrative analysis of a small database, constructed from the results of three MTMM-studies.

4. An illustration: the effects of response scale, order and position

4.1 The data

In 1988, three MTMM-studies were carried out in the Netherlands. The three datasets were collected from the tele-interview panel of the NIPO (a dutch Gallup organisation). A tele-interview panel is a representative panel of respondents that is provided with

home computers. The respondents can use these computers as they want, and in return they have to answer computerised interviews regularly. For dataset 1, respondents were asked to give their opinion about 3 changes in the board of the Dutch labour union that had taken place at that time:

- The chairman Kok switched from labour union chairmanship to national politics
- The chairman Pont, who succeeded Kok, switched from labour union chairmanship to the department of home affairs
- The new chairman, that succeeded Pont, became Stekelenburg

For dataset 2 and dataset 3, respondents were asked how interesting they judged the television broadcastings of 3 olympic wintergames in Calgary. For dataset 2, these games were:

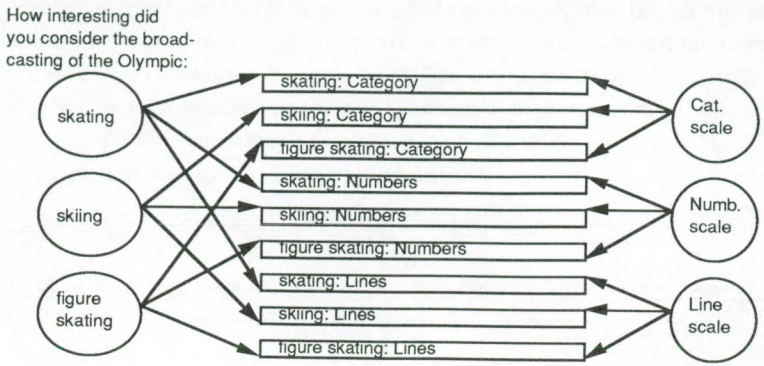
- Skating
- Skiing
- Figure skating

For dataset 3, these games were:

- Skiing
- Ice hockey
- Ramp jumping

For every dataset, the topics were presented 3 times to the same respondents: each time the same question was asked: "What do you think of this event, is it very good or very bad?" (dataset 1) or "how interesting did you consider this television broadcast?" (datasets 2 and 3). In each presentation however, a different response scale was used. Once the respondents had to answer on a category scale consisting of 7 labelled categories; once they had to indicate their opinion by giving a number between 1 and 1000; and once they had to draw a line by moving the cursor on their screen from 1 up to maximally 38 positions, relative to the magnitude of their answer. As an illustration, the measurement model used for dataset 2 is given in figure 1.

figure 1
measurement model for dataset 2: television broadcasts of Olympic winter games



In each of the three studies, respondents were randomly assigned to 1 of 4 possible conditions. These conditions differed in the order in which the different methods were presented: some respondents had to answer on the category scale first, and then on the other two scales, while for others the first scale they had to use was the number production scale or the line drawing scale. In addition, the conditions differed in the position of the questions within the total questionnaire: for some respondents, the described topics were the first topics in the questionnaire, for other respondents, the topics were presented at the end of the questionnaire. The four conditions, and the number of subjects per condition in each dataset are presented schematically in table 1. The total questionnaire consisted of questions about several topics, and lasted about 20 minutes.

table 1
experimental conditions and number of subjects per condition

Condition	First topic in questionnaire?	Which method first?	N dataset 1	N dataset 2	N dataset 3
1	yes	catg.	438	199	230
2	yes	numb.	398	-	-
2	yes	line	-	221	265
3	no	catg.	416	205	234
4	no	numb.	403	-	-
4	no	line	-	226	251

4.2 The MTMM-analyses

For each of the three datasets, we had a 3 (traits) x 3 (methods) MTMM design (see figure 1), for 4 groups: in conditions 1, 2, 3 and 4. A MTMM-true score analysis was carried out for each of the groups in every dataset, to estimate all the validity coefficients, method effects, and reliability coefficients. These analyses were done with the LISREL 7 program. To clear the nature of the quality estimates that resulted from the MTMM-analyses, a table with these estimates obtained from one of the datasets is presented here (table 2). The results obtained for the other two datasets had exactly the same structure and comparable chi squares.

table 2

quality estimates, derived for dataset 2 with the true score model, for each condition

condition 1	validity			method effects			reliability		
methods	cat.	numb.	line	cat.	numb.	line	cat.	numb.	line
traits									
1 Skating	.95	.98	.99	.32	.21	.12	.93	.96	.98
2 Skiing	.95	.98	.99	.32	.21	.12	.93	.94	.97
3 Figure skating	.95	.98	.99	.32	.21	.12	.92	.96	.98
$\chi^2 = 48.74$, df = 21									
condition 2	validity			method effects			reliability		
methods	cat.	numb.	line	cat.	numb.	line	cat.	numb.	line
traits									
1 Skating	.94	.98	.98	.33	.18	.22	.88	.98	.95
2 Skiing	.94	.98	.98	.33	.19	.22	.88	.95	.94
3 Figure skating	.94	.98	.98	.33	.18	.21	.89	.98	.98
$\chi^2 = 37.38$, df = 21									
condition 3	validity			method effects			reliability		
methods	cat.	numb.	line	cat.	numb.	line	cat.	numb.	line
traits									
1 Skating	.97	.97	1.00	.23	.24	.02	.93	.97	.91
2 Skiing	.97	.97	1.00	.25	.24	.02	.85	.97	.91
3 Figure skating	.97	.97	1.00	.25	.24	.02	.86	1.00	.94
$\chi^2 = 19.84$, df = 21									
condition 4	validity			method effects			reliability		
methods	cat.	numb.	line	cat.	numb.	line	cat.	numb.	line
traits									
1 Skating	.96	.97	.98	.27	.23	.20	.88	.96	.94
2 Skiing	.96	.97	.98	.28	.25	.21	.84	.92	.91
3 Figure skating	.97	.97	.98	.26	.23	.20	.92	.98	.94
$\chi^2 = 40.80$, df = 21									
mean		.97			.22			.93	
standard deviation		.02			.09			.41	

We can infer some effects by just looking at this table: the method effects are all rather small, but the highest method effects are found for the category scale, which also generally has the lowest validity. The reliability coefficients show more variation between the traits than the validity and method effect coefficients. However, if we want to compare the quality estimates between the conditions, methods, and traits at the same time, it becomes rather complicated to infer anything from this table. Besides that, we can not easily compare the magnitudes of the differences. Therefore, a small scale meta-analysis was carried out, to test wether the variance in the obtained quality estimates could be explained by the different traits, by the methods used, or by the different conditions.

4.3 The Multiple Classification Analysis

The coefficients and characteristics of the questions from all three datasets were put together into one database. In table 3, the coefficients from dataset 2 are presented again as an illustration of the structure of such a database: the same coefficients can be found in table 2, but now they are accompanied by codes to identify for which item and in which condition they were obtained.

table 3
structure of the data for meta-analysis

coefficients			characteristics/factors				
val.	meth.	rel.	first topic	order	method	trait	dataset
.95	.32	.93	1*	1	1	1	2
.98	.21	.96	1	1	2	1	2
.99	.12	.98	1	1	3	1	2
.95	.32	.93	1	1	1	2	2
.98	.21	.94	1	1	2	2	2
.99	.12	.97	1	1	3	2	2
.95	.32	.92	1	1	1	3	2
.98	.21	.96	1	1	2	3	2
.99	.12	.98	1	1	3	3	2
.94	.33	.88	1	2	1	1	2
.98	.18	.98	1	2	2	1	2
.98	.22	.95	1	2	3	1	2
.							
.							
etc.							

*0= no
1= yes

Next, Multiple Classification Analyses could be carried out in which the factors were: trait, method, order of presentation and position in the questionnaire (first topic or not) and the dependent variables were the validity coefficients or the reliability coefficients. For the actual analyses, the LISREL estimates of these coefficients were multiplicated by 100. No tests were done on the method effect coefficients because in the true score

model, the method variance is the complement of the valid variance, so the method effects do not provide any new information, once the validity is known. Multiple Classification Analysis is available in SPSS, as an option of the analysis of variance procedure (ANOVA). It provides the usual analysis of variance tables and in addition tables of category means for each factor, expressed as deviations from the grand mean. In table 4 and table 5 the analysis of variance results are presented.

table 4
ANOVA on **validity** coefficients

Source of variation	SS	df	MS	F	Signif.
FIRST*	.59	1	.59	.58	>.05
ORDER	29.17	2	14.59	14.22	.00
TRAIT SUBJECT	.38	1	.38	.37	>.05
METHOD	134.06	2	67.03	65.32	.00
explained	207.03	6	34.51	33.63	.00
residual	103.64	101	1.03		
total	310.67	107	2.90		

*"First" in this and following tables stands for: First topic in questionnaire or not, "Order" stands for: order of presentaton of the methods (a method could be presented as the first, second or third method in the questionnaire). "Trait subject" indicates wether the subject of the traits was the labour union or the Olympic games.

table 5
ANOVA on **reliability** coefficients

Source of variation	SS	df	MS	F	Signif.
FIRST	.59	1	.59	.08	>.05
ORDER	208.90	2	104.45	14.44	.00
TRAIT SUBJECT	67.78	1	67.78	9.37	.00
METHOD	80.06	2	40.03	5.53	.00
explained	462.46	6	77.08	10.65	.00
residual	730.73	101	7.24		
total	1193.19	107	11.15		

When interpreting the ANOVA tables, it should be taken into account that the homogeneity of variance assumption is not met and that there are some complications in the data, like dependance between the LISREL quality estimates and correlations between predictors. The significance tests may therefore not be justified, but the tables still give a global indication of the effects: if we look at the sum of squares of each factor relative to the total sum of square, we see that order and method seem to have most effect on the validity coefficients; and order, method and trait subject on the reliability coefficients. In the Multiple Classification table (table 6) the effects are shown more specified. This table also clears the effect of the correlations between [illegible]tors

by showing the effects of each predictor on the quality estimates, both before and after adjusting for the effects of all other predictors.

table 6
multiple classification analyses: effects of factor levels on quality estimates

factors and levels	N number of estimates	Validity Unadjusted Dev'n	Validity Unadjusted Eta^2	Validity Adjusted Dev'n	Validity Adjusted $Beta^2$	Reliability Unadjusted Dev'n	Reliability Unadjusted Eta^2	Reliability Adjusted Dev'n	Reliability Adjusted $Beta^2$
FIRST									
no	54	-.07		-.07		-.07		-.07	
yes	54	.07		.07		.07		.07	
			.00		.00		.00		.00
ORDER									
first	36	-1.00		-.75		-1.92		-1.58	
second	36	1.00		.53		2.23		1.93	
third	36	.00		.22		-.32		-.35	
			.23		.10		.26		.19
TRAIT SUBJECT									
Labour Union	36	.08		.08		1.12		1.12	
Olympic games	72	-.04		-.04		-.56		-.56	
			.00		.00		.06		.06
METHOD									
Category	36	-1.67		-1.45		-1.85		1.27	
Numbers	36	.22		.06		.93		.72	
Lines	36	1.44		1.39		.93		.55	
			.56		.48		.15		.07
R^2 ADJUSTED (Joint explanatory power of the factors)					.67				.39

table 7
mean validity and reliability coefficients

	Validity	Reliability
Mean	97.11*	94.63*
Standard deviation	1.70	3.34

In the Multiple Classification table, the effects are expressed as deviations from the grand mean. These deviations show the category effects for each factor, that is: how much the validity or reliability would go up or down from the mean (presented in table 7) if a measure had that specific characteristic. The unadjusted categories are presented as well as the category effects adjusted for the effects of all other factors. This means,

* Note that the estimates of validity and reliability are based on the LISREL parameters x 100. These estimates range between 84 and 100.

for example, that the total mean validity of 97.11 is lowered by 1.45 when a category scale is used, is increased with .06 when a number estimation scale is used, and is increased with 1.39 when a line drawing scale is used, holding everything else constant. So the mean validity coefficients for the three different methods are (when the effects of the other factors are taken into account).

table 8
example: the effects of method

	mean validity coefficient
METHOD	
category	95.66
numbers	97.17
lines	98.50

The differences between the unadjusted and adjusted deviations are an indication of the amount of intercorrelation the analysis adjusted for. The squared betas that are also given in table 6 indicate the adjusted importance of a factor in accounting for the variance in the validity and reliability coefficients, that is: when all other factors are held constant. The (unsquared) betas are equivalent to standardized partial regressions coefficients in the sense used in multiple regression. The eta squared that can be found in the column "Unadjusted", indicates the proportion of variance explained by a factor when considered alone. The etas are equivalent to correlation ratios. As can be inferred from these measures, the validity is influenced mostly by the factor method, both when considered alone and in combination with all other factors. The factor "order of presentation" also has a moderate effect on validity: the mean validity is lowest when a method is the first one presented and is highest when it is the second, everything else being constant. The factors first and trait have almost no effects on the validity. For the reliability, some effects of the factor order of presentation were found: the mean reliability of 94.63 is 1.58 *lower* when a method is the presented first and 1.93 *higher* when a method is presented as second.

Some extra Multiple Classification Analyses were carried out to test wether any first-order interaction effects accounted for a substantial part of the variance. This was done by combining two variables into one new variable, like is shown in table 8, and running a separate analysis for each combination variable. The effects of a combination variable include both the main effects and interaction effects of the original variables. The explanatory power of the combination variable can then be compared with that of the additive multiple classification results. It turned out that introducing a method x order factor increased the explained variance in the reliability coefficients somewhat: the effects of this factor are presented in table 9. No substantial interaction effects were found for the validity coefficients.

table 9
multiple classification analyses: effects of method x order on reliability estimates

factor levels	N number of estimates	Reliability Unadjusted Dev'n	Eta^2	Adjusted* Dev'n	Beta^2
METHOD*ORDER					
Category					
first	18	-2.02		-2.02	
second	6	3.70		3.06	
third	12	-4.38		-4.06	
Numbers					
first	6	-2.63		-3.28	
second	12	1.70		2.03	
third	18	1.59		1.59	
Lines					
first	12	-1.38		-1.06	
second	18	2.09		2.09	
third	6	2.04		1.39	
			.53		.49
R^2 ADJUSTED (Joint explanatory power of the factors)					.54

* The other factors in this analysis, for which the effects of method x order are adjusted, were again TRAIT SUBJECT and FIRST. The effects of these factors remained the same as in table 5 and are therefore not shown here.

From this table it can be concluded that, although the reliability of a measure is generally lower when it is the first method presented, this effect is not as strong for the line drawing scale as it is for the other two scales. The category scale also decreases reliability when presented as third method.

4.4 The effects of response scale, order and position

The results of this study suggest that a line drawing scale might be the most valid method to measure opinions about political events or television broadcasts. This conclusion is in agreement with findings from some other studies that have been done in the context of the same international measurement project (Ohlsson and Roe,1990). The 7 points verbal category scale turned out to be inferior to both the number production scale and line drawing scale in the present study: it produced considerably lower validity estimates and its reliability was more dependent upon the order of presentation than the reliability of the other scales. The fact that some other factors included in this study did *not* affect the validity or reliability coefficients is of interest as well: apparently the quality of these data is not influenced much by trait content or by the position of the topics in the questionaire.

5. Applications

The above conclusions are based on a small number of datasets and an analysis with only four factors. How general the effects of line drawing scales, category scales, topics and position really are, will be explored in the international measurement project of which this study is a part, together with the effects of many more characteristics. In the final large scale database there will be much more variation in quality estimates then in the illustrative study described here, because this study consisted of three quite similar datasets collected within one country. The effects of the factors in the meta-analysis will therefore probably be more pronounced then in the present study. With the information provided by the final meta-analysis, it will be possible to determine the expected level of validity for a specific survey question on the basis of its characteristics, and to decide for example, to improve this validity by using another response scale or by putting the item somewhere else in the questionnaire. It will also be possible, as Andrews (1984) has shown, to correct observed correlations between variables on the basis of the information about the measurement quality, to derive the true relationship between the variables of interest. It can be shown, using path analysis, that:

$$\rho(y_1\ y_2) = \lambda^t_1 * \rho(\xi_1\ \xi_2) * \lambda^t_2 + \lambda^m_1 * \lambda^m_2 \qquad (9)$$

where: $\rho(y_1\ y_2)$ is the observed correlation between two measures that each tap a different trait.
$\rho(\xi_1\ \xi_2)$ is the true correlation between the traits tapped by y_1 and y_2.
λ^t_1 is the validity coefficient of measure 1
λ^t_2 is the validity coefficient of measure 2
λ^m_1 is the method effect of measure 1
λ^m_2 is the method effect of measure 2

This formula can be transformed to provide predictions of the true relationship:

$$\rho(\xi_1\ \xi_2) = [\rho(y_1\ y_2) - \lambda^m_1 * \lambda^m_2] / [\lambda^t_1 * \lambda^t_2] \qquad (10)$$

This equation shows that it is necessary to have estimates of validity and method effects if one is interested in correlations between concepts. The approach that is followed in the international measurement project will provide such estimates for a large number of different types of survey measures. In addition, it will give information about the comparability of survey results between countries.

References

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

Andrews, F.M. Morgan, J.N., Sonquist, J.A., & Klem, L. (1973). *Multiple classification analysis.* Ann Arbor, MI: Institute for Social Research.

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

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multimethod-multitrait matrix. *Psychological Bulletin , 56,* 833-853.

Joreskog, K.G. & Sorbom, D. (1983) *Lisrel VI: A general computer program for estimation of linear structural equation systems by maximum likelihood methods.* Uppsala: Department of Statistics.

Ohlsson, A. & Roe, K. (1990) Measurement validity and sensitivity: an analysis of line production and number estimation scales. *Working papers from the research project "Measuring Measurement",* University of Göteborg.

Saris, W.E. & Andrews, F.M. (1991). Evaluation of measurement instruments using a structural modeling approach. *Forthcoming.*

Saris, W.E. & van Meurs, A. (1990). *Evaluation of measurement instruments by meta-analysis of Multitrait-multimethod studies.* Amsterdam: North Holland.

ontvangen 13- 2 -1991
geaccepteerd 15- 10-1991

Appendix A

Pathdiagrams of the models

figure 1

the standard model

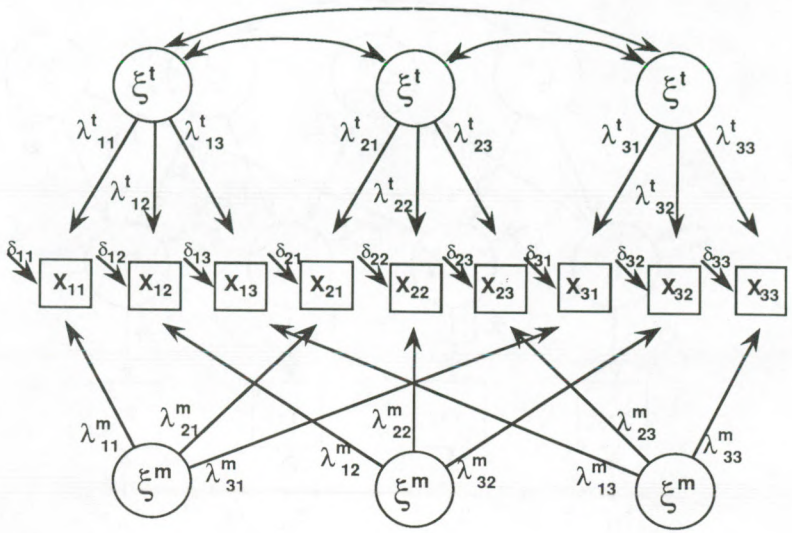


figure 2

true score model