PROBLEMS IN RUSSIAN OPINION POLLS
Memory Effects, Acquiescence, and Awareness

William M. van der Veld, Willem E. Saris

University of Amsterdam

Abstract

Many researchers, beginning with Campbell and Fiske (1959), Andrews (1984), and Saris and Andrews (1991), have suggested to use the multitrait multimethod approach to estimate the quality of measurement instruments. This approach has also been used in Russia, building on the experience of Andrews (1984) in the USA, Költringer (1995) in Austria, and Scherpenzeel and Saris (1997) in The Netherlands. Analyses of these data sets revealed no insurmountable problems, whereas analysis of the Russian data by this approach has raised serious problems with many of the experiments. In particular, the multitrait multimethod experiments failed with topics where the correlation between repeated observations of the same trait within the interview is extremely high compared with the correlation between repeated observations of the same trait across the interviews. This indicates a contradiction, i.e. the high correlation within the interviews indicates a stable opinion, whereas the low correlation across the interviews indicates a very unstable opinion. Assuming that the cross-interview correlation gives a more realistic picture, we argue that the high correlation within the interviews are contaminated by memory effects and/or response set effects, such as acquiescence. Although this is the first time that we have found these effects in multitrait multimethod experiments they may be more general than so far realised.

Correspondence to: Drs. W. M. van der Veld ing. (vdveld@worldonline.nl)

University of Amsterdam - Department of Communication Sciences (ASCoR)

Oude Hoogstraat 24, 1012 CE Amsterdam.

Note: the above e-mail address can also be used to ask for the survey questions and the correlation matrices of the topics dealt with in this paper.

1. AN INTRODUCTION TO THE MULTITRAIT MULTIMETHOD DESIGN

Campbell and Fiske (1959) suggested that the quality of a measurement instrument, viz. the question and measurement procedure, could only be determined by comparing it with other instruments. Such comparisons were made by inspecting a specific correlation matrix (the multitrait multimethod matrix) that is obtained from repeated measures of a set of traits. Each set of traits is measured using a different measurement procedure (method). Different measurement procedures can be produced by any difference, such as different data collection procedures, different place of the question in the survey, different response scale lengths, differences in response scale labels, the presence or absence of a no opinion filter, et cetera.

An example of a multitrait multimethod matrix, obtained from a set of three questions concerning political efficacy, is presented in table 1. The political efficacy measure consists of the following three statements:

R_1 *I do not think public officials care much what people like me think.*

R_2 *People like me do not have anything to say about what the government does.*

R_3 *Sometimes politics seems so complicated that a person like me cannot really understand what is going on.*

These statements were administered twice within an interview, and the interviews were repeated with the same respondents after an interval of one year. The respondents were asked to express their agreement with each set of statements in four different ways. The first time in the first interview, respondents were asked to express their opinion on a 7-point disagree-agree scale; the second time they had to express their opinion on a 5-point agree-disagree scale. The first time in the second interview respondents were asked to express their opinion on a 6-point agree-disagree scale, and the second time they were asked to express their opinion by drawing a line (the longer the line, the more they agreed with the statement). The combination of three questions (traits) and four measurement procedures (methods) leads to 12 observed variables, which are presented in table 1.

TABLE 1: A multitrait multimethod matrix of Pearson correlation coefficients obtained from a combination of three questions on political efficacy and four measurement procedures[i].

	First interview						Second interview					
	7p agree-disagree			5p agree-disagree			6p agree-disagree			line production		
	R_{111}	R_{211}	R_{311}	R_{112}	R_{212}	R_{312}	R_{121}	R_{221}	R_{321}	R_{122}	R_{222}	R_{322}
R_{111}	1.00											
R_{211}	0.27	1.00										
R_{311}	0.19	0.49	1.00									
R_{112}	-0.58	-0.23	-0.16	1.00								
R_{212}	-0.26	-0.76	-0.50	0.30	1.00							
R_{312}	-0.17	-0.47	-0.81	0.18	0.56	1.00						
R_{121}	-0.12	-0.08	-0.05	0.13	0.09	0.06	1.00					
R_{221}	-0.09	-0.25	-0.23	0.09	0.27	0.24	0.37	1.00				
R_{321}	-0.07	-0.22	-0.34	0.05	0.23	0.33	0.27	0.53	1.00			
R_{122}	0.12	0.09	0.07	-0.11	-0.09	-0.08	-0.34	-0.20	-0.15	1.00		
R_{222}	0.09	0.25	0.23	-0.06	-0.26	-0.26	-0.18	-0.50	-0.41	0.36	1.00	
R_{322}	0.09	0.24	0.35	-0.04	-0.24	-0.35	-0.16	-0.38	-0.60	0.26	0.59	1.00

There is a whole family of multitrait multimethod models that can be used to test the possible effects of measurement procedures on the correlations between the observed variables (Coenders and Saris, forthcoming). We will review two frequently used models.

The first is the true score multitrait multimethod model (Saris and Andrews, 1991). This model has recently been used in several studies to determine the effect of question characteristics on the quality of questions, particularly their reliability and validity. In The Netherlands, Scherpenzeel and Saris (1997), and in Austria Költringer (1995) conducted studies on the quality of questions across different topics. Furthermore, Scherpenzeel and Saris (1995) carried out a study of questions on the quality of life-satisfaction across ten different European countries.

[i] In the correlation matrix the indices in the variable names refer to a more general use of variable indices throughout this text, such that each variable is indicated by R_{ijk}, where the i refers to a specific variable, j to the interview in which this variable was observed, and k to the first, second or third observation of variable i in interview j.

The second model is the correlated uniqueness model (Marsh and Bailey, 1991) which is a less restrictive version of many other multitrait multimethod models. It is therefore a very useful model for testing different hypotheses about the additive (Andrews, 1984) or multiplicative (Browne, 1984) effects of the applied measurement procedures on the structure on the covariance matrix.

1.1. The true score multitrait multimethod model

The true score multitrait multimethod model (Saris and Andrews, 1991) is specified as:

$$R_{ijk} = \lambda_{ijk}\, \tau_{ijk} + \varepsilon_{ijk} \qquad \text{for all i, j, and k.} \qquad (1)$$

$$\tau_{ijk} = \gamma_{ij}\, \tau_{ij} + \gamma_{ik}\, F_k + \zeta_{ij} \qquad \text{for all i, j and k.} \qquad (2)$$

where

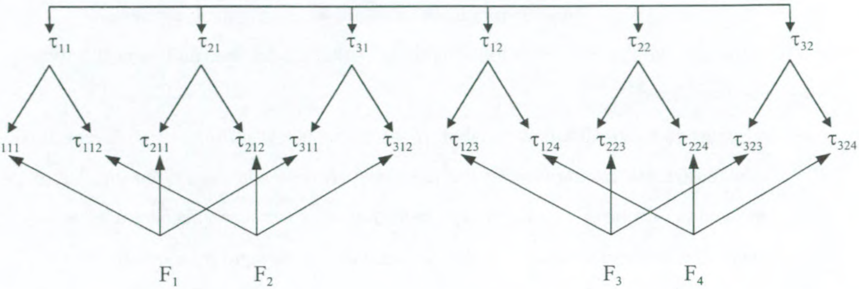
- R_{ijk} is the standardised observed variable of trait i measured in wave j with the kth measurement procedure;
- τ_{ijk} is equal to R_{ijk} corrected for the random measurement error;
- τ_{ij} is a latent factor representing trait i in wave j (trait factor);
- F_k is the latent factor due to the measurement procedure (method factor);
- ε_{ijk} is the random measurement error in the observed score (random error term); and
- ζ_{ij} is the disturbance term in the equation for τ_{ijk}.

The first equation defines the relation between a standardised observed variable and the same variable after correction for random measurement error, called the true score in psychometric literature (Lord and Novick, 1968). The λ_{ijk} coefficient is a standardised coefficient indicating the strength of the relation between those two variables. The second equation suggests that the true score is affected by the trait one would like to measure, the applied measurement procedure (due to the way people answer or react to it), and a disturbance term. The second equation of the specified model is presented in Figure 1.

In this model, it is assumed that the random error terms are not correlated with the method factors, the trait factors, or the disturbance terms. The trait factors are all correlated. This model explicitly allows for change of opinions across the interviews.
In order to improve the empirical identifiability of this model, normally the following assumptions are made.

- All method factors are uncorrelated. If the methods are sufficiently different, this is a reasonable assumption. If the measurement procedures are too similar, respondents' reactions will be pretty much the same for such measurement procedures, and correlation will be found between the corresponding method factors.
- All disturbance terms are zero. This assumption can be met by keeping the formulation of a question unchanged throughout repeated administrations of a question, regardless of differences due to the measurement procedures. If this is done one can be certain that the same trait is being measured apart from random measurement error (Saris, 1982).
- All random error terms are uncorrelated with each other, meaning that the correlation between the (observed) variables is solely due to the trait(s) of interest and the measurement procedure(s) used. This assumption will be violated if other factors contaminate the relationship between two observations. Memory effects and response set effects are commonly considered as such. In the model, response set effects between a set of different variables observed with the same measurement procedure are accounted for by the method factors. Response set effects across repeated observations of the same variable are assumed to be absent. Memory effects are also assumed to play no part because the time between the repeated observations is always at least twenty minutes (Van Meurs and Saris, 1995).

FIGURE 1: A two-wave multitrait multimethod panel model with three traits and four methods, allowing for change over time of the traits.



The effects of the different variables on each other are interpreted as follows.

- λ_{ij} is the reliability coefficient. The square of this parameter can be interpreted as an estimate of the test-retest reliability (Heise and Bohrnstedt, 1970; Lord and Novick, 1968).
- γ_{ij} is the true score validity coefficient. The square of this coefficient is the explained variance by the trait of interest.
- γ_{ik} is the method effect. The square of this coefficient is the explained invalid variance due to the applied measurement procedure.

1.2. The correlated uniqueness model

The correlated uniqueness model (Kenny, 1976; Marsh, 1989; Marsh and Bailey, 1991) is a less restrictive version of many other multitrait multimethod models. This model allows tests for the type of effect, i.e. additive or multiplicative, of the measurement procedure on the structure of the covariance matrix. Hence, this model can be used generally for carrying out multitrait multimethod analyses (Coenders and Saris, forthcoming). Furthermore, the correlated uniqueness model is very stable in that it hardly ever leads to problems of empirical underidentification, failure to converge, or inadmissible estimates (Marsh, 1989; Marsh and Bailey, 1991).

The correlated uniqueness model is specified as:

$$R_{ijk} = \lambda_{ijk}\, \tau_{ij} + \delta_{ijk} \qquad \text{for all i, j, and k} \qquad (3)$$

where

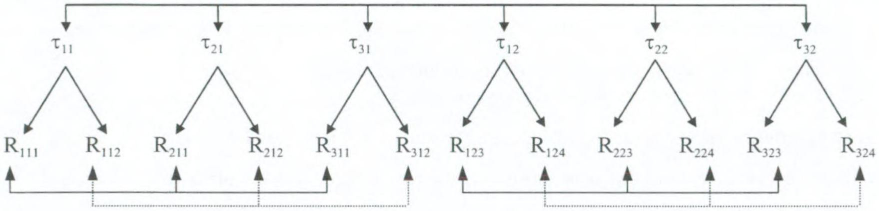
- R_{ijk} is the observed variable of trait i measured in wave j with the k^{th} measurement procedure;
- τ_{ij} is a latent factor representing trait i in wave j (trait factor); and
- δ_{ijk} is the random measurement error in the observed score (random error term).

Equation 3 defines the relation between the observed variable and the trait of interest. The parameter λ_{ijk} is a standardised coefficient indicating the strength of the relation between the observed variable and the trait of interest. This parameter can be interpreted as an indicator of the quality, and is under certain conditions equal to the product of the λ_{ijk} and γ_{ij} from the true score multitrait multimethod model.

In this model, it is assumed that the error terms are not correlated with the trait factors. It is also assumed that the error terms are not correlated with each other ($\theta_{\delta ij}$=0), except for the covariance between error terms of variables observed with the same measurement procedure. This model implies that the correlation between the (observed) variables is solely due to the trait(s) of interest and the measurement procedure(s) used. This assumption will be violated if other factors contaminate the relationship between two observations, such as memory effects or response set effects.

In the correlated uniqueness model, the covariances between the error terms of variables observed with the same measurement procedure are allowed to differ for each pair of observed variables. In that case the effect of the measurement procedure is neither specified as additive nor as multiplicative. If however, the covariances between the error terms of variables observed with the same measurement procedure are constrained to be equal for each measurement procedure, then the correlated uniqueness model is locally equivalent to the true score multitrait multimethod model; under the condition that both models have the same number of free parameters, and the estimated parameters of either model are not inadmissible (Luijben, 1989). If the covariances between the error terms of variables observed with the same measurement procedure are constrained to be equal to the covariance between the traits of interest multiplied by a constant, which is specific for the measurement procedure involved, then the correlated uniqueness model is said to be multiplicative and locally equivalent to the direct product model (Browne, 1984); under the condition that there are three traits and three methods.

FIGURE 2: A two-wave correlated uniqueness panel model with three traits and four methods, allowing for change over time of the traits.



2. ANALYSIS OF THE DATA

Using the true score multitrait multimethod model, Scherpenzeel and Saris (1995, 1997), and Költringer (1995), have studied, using LISREL, many topics and evaluated data quality across different countries. They evaluated the data quality for many characteristics of measurement instruments (survey questions) with a meta-analysis. These characteristics included domain of the question (politics, values, etc.), social desirability of the subject of the question, length of the question, type of response scale used, the employed mode of data collection, the context of the question, and many more. In this meta-analysis, the effect of each characteristic on validity and reliability is estimated using regression analysis. When the size of these effects are known, one can predict the quality of the data produced by any measurement instrument in advance of a study, enabling the researcher to improve the quality of his measurement instruments. In addition, these quality predictions can be used tentatively to correct for measurement error in analyses of any data set obtained by questionnaire surveys with any data collection method (Saris, Van der Zouwen, 1999; Saris, Van der Veld, 2000). We planned to follow the same path in Russia as did Scherpenzeel and Saris for several European countries.

The data were collected in the RUSSET[i] panel by the Institute for Comparative Social Research (CESSI) in Russia. The panel has started in 1993 and stopped in 1999. The respondents were selected by a two-stage sampling procedure, where in the first stage 50 areas were drawn and in the second stage a random sample was drawn proportional to the size of the area. In addition, in order to reduce panel dropout, respondents were allowed to skip an interview and enter the panel again the following year. The respondents were interviewed face-to-face using the paper and pencil method. Only few interviewers revisited the same respondent again. Furthermore, inspection of the birth-dates of each panel-member showed that it is very unlikely that interviewers filled in the questionnaire themselves. Appendix 1 gives the sample size and aspects of the data preparation. There are 11 topics included in the multitrait multimethod experiments, leading to a total of 200 measurement instruments to be evaluated for quality.

2.1. Results of the multitrait multimethod analyses

In the analyses of the data with the true score multitrait multimethod (MTMM) model serious problems were encountered with respect to non-convergence and unacceptable estimates. Therefore, the data were also analysed with the correlated uniqueness (CU) model. The analyses of the data with both models will be discussed in the next two sections.

2.1.1. Analyses with the true score multitrait multimethod model

The poor results of the analyses of the data with this model are presented in table 2. Only 5 out of 14 analyses yielded acceptable results, nevertheless they fitted poorly. In four data sets the model yielded unacceptable estimates due to negative variances. Furthermore, in four other cases the program did not converge to a solution.

TABLE 2: The results of the analyses of the data sets using the true score multitrait multimethod model.

Topic	Chi-square	df	Remarks
Buying (1)	845.9	35	Unacceptable estimates
Buying (2)	450.0	35	Unacceptable estimates
Change	116.1	35	Bad fit, acceptable estimates
Ingroup	140.4	35	Bad fit, acceptable estimates
Nationality (1)	-	-	Did not converge
Nationality (2)	-	-	Did not converge
Outgroup	-	-	Did not converge
Political efficacy	68.0	35	Poor fit, acceptable estimates
Policy	347.5	72	Unacceptable estimates
Satisfaction (1)	147.7	45	Bad fit, acceptable estimates
Satisfaction (2)	50.5	35	Unacceptable estimates
Spending	342.7	72	Bad fit, acceptable estimates
Threat	657.0	72	Very bad fit
Trust	-	-	Did not converge

Anderson and Gerbing (1984) and Rindskopf (1984) have studied these problems (negative variances and non-convergence) and recommend sample sizes of 150 cases or more and at least two indicators per factor. In our data sets, the sample sizes are always considerably larger (Appendix A) than the size recommended by these authors. On the other hand, we have only measured two indicators per trait factor in each interview. Nevertheless, the trait factors are correlated with many other trait factors, which contributes considerably to the empirical identifiability of the model. It is therefore unlikely that the number of indicators per factor is the cause of the problem in this case. It is more likely that the problems are due to incorrect specification of the effects in this model. For example, in this model the method effects are specified as additive but perhaps they should be specified as multiplicative (Coenders and Saris, forthcoming), or perhaps the method effects are neither additive nor multiplicative, and therefore the unrestricted correlated uniqueness model should be fitted to these data. We will test these hypotheses in the following analyses using the correlated uniqueness model.

[i] Russian Socio-economic Transition panel. This study was initiated by Willem E. Saris and is supported by the Dutch organisation for Scientific Research (NWO).

2.1.2. Analyses with the correlated uniqueness model

Since, the additive specification of the correlated uniqueness model is equivalent to the true score multitrait multimethod model (in our specific case), it is unnecessary to test that specification of the correlated uniqueness model. So, we have analysed only topics that yielded unacceptable results in the analyses with the true score multitrait multimethod model. The results of the analysis with the unrestricted correlated uniqueness model, presented in table 3, are again very poor. Only 2 out of 9 analyses produced acceptable results. Furthermore, the program did not converge to a solution for 'buying', 'policy', 'nationality', and 'threat'.

These results are peculiar. The correlated uniqueness model rarely leads to problems of empirical underidentification[i], failure to converge, or unacceptable results (Marsh and Bailey, 1991), but our results do not confirm this. Since the unrestricted correlated uniqueness model only produced acceptable result for two topics, no further attempt was made to analyse these topics using the correlated uniqueness model with multiplicative restrictions.

TABLE 3: The results of the analyses of the data sets using the unrestricted correlated uniqueness model.

Topic	Chi-square	df	Remarks
Buying (1)	65.0	27	Poor fit, acceptable estimates
Buying (2)			Did not converge
Nationality (1)	-	-	Did not converge
Nationality (2)	-	-	Did not converge
Outgroup	72.5	52	Unacceptable estimates
Policy	-	-	Did not converge
Satisfaction (2)	28.5	27	Good fit, acceptable estimates
Threat	-	-	Did not converge
Trust	87.1	52	Unacceptable estimates

[i] We found that for topics were the across wave correlations were close to zero for one or more variables, the model suffered from problems of underidentification.

In short, it seems that both the true score multitrait multimethod model and the correlated uniqueness model are unsuitable for these data. Although both models produced acceptable results for some topics, in most cases the models in their standard form produced unsatisfactory results. Andrews (1984), Költringer (1995), and Scherpenzeel and Saris (1997) used the same multitrait multimethod design and analysed the data successfully with the true score multitrait multimethod model. They also encountered problems, but these could be overcome by minor adjustments. In order to investigate whether the Russian data differ in some way from the Dutch data, we will inspect the correlation coefficients on comparable topics for both countries.

3. A MORE DETAILED LOOK AT THE PROBLEMS

Looking at those topics that gave an acceptable solution and those that did not, the impression was given that respondents' awareness of the topic might play an important role. For example, the topic 'satisfaction' and 'political efficacy' resulted in reasonable solutions but topics like 'nationality', 'outgroup', and 'threat' did not.

In order to find out why the Dutch data (Scherpenzeel, 1995; Scherpenzeel and Saris, 1997) did not produce so many problems as the Russian data, we have compared them. We selected topics for the comparison based on two criteria; firstly that the topics should be comparable across the two countries, and secondly that the topics should be comparable with respect to the degree of respondents' awareness of the topic.

Many investigators have found that awareness is an important factor in opinion research, e.g. Converse (1964), Schuman and Presser (1981), Billiet et al (1986), Lodge et al (1995), Luskin (1997), Sniderman et al (1991), and Zaller (1992). Billiet et al. and Schuman and Presser have argued that awareness will lead to crystallisation of an opinion. This can be taken to mean that, in the case of a more crystallised opinion the number of responses on a scale with possible answers acceptable for a respondent is smaller. This would mean that repeated observations in a short time should show a high similarity in responses for people that have a crystallised opinion. Crystallisation will also lead to a more stable opinion over a longer period of time because the opinion will be less affected by new information than a less crystallised opinion. In short, the more crystallised an opinion is, the higher the correlation will be across repeated observations of that opinion. We will use this notion throughout the rest of this text.

Three topics were selected on the basis of different degrees of awareness. The first topic was satisfaction with different aspects of the respondent's life, which is available for both countries. The respondents were asked to evaluate their satisfaction with their financial situation, their housing conditions, their social contacts, and their life in general. We assume that all people can evaluate this topic rather easily because everybody is aware of these aspects, which should lead to a stable opinion. The attention paid by politicians to the public, as measured by the 'political efficacy' questions, is the second topic. This topic, which is also available for both countries, is already a step farther away from people's everyday experience. Nevertheless, people still will gain an impression of this via the different mass media. We would expect that the stability of opinion on these issues should be lower than on satisfaction. Moreover, we expect that in Russia the stability is even lower than in The Netherlands because of ongoing political changes in Russia. Thirdly, because no single topic was available at this level of awareness for both countries, we selected two topics that are approximately equal with respect to degree of awareness. Thus, for Russia the third topic is the problem of 'nationality', i.e. the development and existence of a Russian nationality and culture. We assume that Russians have a very low level of awareness of this topic after the collapse of the USSR and the transition to a new political unity. For The Netherlands, the third topic is 'European unification'. We assume that the Dutch have a very low level of awareness of this topic, since it is too distant from people's everyday experience.

The correlation between the same variables on comparable topics for Russia and The Netherlands are presented in table 4. All data originate from multitrait multimethod multi-time design, such that each variable is measured twice within an interview and two interviews are administered with the same respondents. In fact, the topics are merely names for a group of questions (variables) that measure different aspects of those topics; such that we can speak of a lowest and highest correlation between the same variables (questions) of a topic.

TABLE 4: Pearson correlation coefficients between the same variables in the Netherlands and Russia.

Correlation between the same variables for	Netherlands		Russia	
	Low	High	Low	High[i]
Satisfaction				
in the first interview	.71	.85	.65	.74
in the second interview	.56	.69	.73	.84
Political efficacy				
in the first interview	.48	.63	.58	.81
in the second interview	.26	.51	.34	.60
European unification/Nationality				
in the first interview	.47	.57	.59	.81
in the second interview	.35	.42	.67	.81
Satisfaction				
across the interviews for the first set of observations	.28	.42	.20	.40
across the interviews for the second observations	.21	.41	.20	.36
Political efficacy				
across the interviews for the first set of observations	.28	.42	.12	.34
across the interviews for the second observations	.22	.42	.11	.35
European unification/Nationality				
across the interviews for the first set of observations	.30	.45	.08	.13
across the interviews for the second observations	.33	.42	.06	.15

First, a remark should be made about both multitrait multimethod studies. In The Netherlands[ii], the time between the administration of the interviews was never longer than three months, whereas in Russia the time between two administrations was approximately a year. Given this difference in the time between the administration of the interviews, comparisons of the across interview correlations, i.e. the stability of the opinion, for both countries may be questionable. One should not therefore conclude that the opinion on a specific topic is more stable in the Netherlands than in Russia. Further-

[i] Low and High refer to the lowest and highest correlation computed for each topic=set of variables.

[ii] In The Netherlands the data were collected by 'Stichting Telepanel'. In their panel, member was provided with a home computer. The respondents then had to download a questionnaire every second week, which was automatically collected also via a phone line.

more, for the repetitions within the interviews, the same rule was applied in both countries, that there should be at least 20 minutes between repetitions in order to prevent (if possible) memory effects (Van Meurs and Saris, 1995).

The following results can be observed in table 4. When the *correlation in the interviews* is compared, we notice a gradual decrease of the correlation in The Netherlands associated with the assumed decrease of respondents' awareness of the topic. In Russia, this pattern of decreasing correlations is not evident. For topics that are assumed to evoke large differences in awareness (satisfaction and nationality) the correlation is very similar. In the Dutch case, when the *correlation across the interviews* is compared, there is hardly any difference in the size of the correlation for topics that differ widely with respect to awareness. In the Russian data, however, a gradual decrease of the correlation is found which is associated with the assumed decrease of the respondents' awareness about a topic.

It seems there are two phenomena that affect the Dutch and the Russian data differently. The first concerns the very large differences in correlations between the same variables within an interview and across the interviews in Russia, especially for topics were we expect a less crystallised opinion. These differences are much smaller in the Dutch data, and for topics where we expect a less crystallised opinion almost absent. The second phenomenon concerns the small differences in correlations between the same variables within an interview in Russia across the topics, whereas we would expect differences due to differences in degrees of awareness. These anticipated differences are clearly evident in the Dutch data. Obviously, if our hypothesis about awareness is true, then in both the Dutch data and the Russian data something is wrong - in the Dutch case relating to the cross-interview correlation and in the Russian case relating to the within interview correlation. Nevertheless, the Dutch data did not lead to serious problems, whereas the Russian data did. The question to be answered therefore is whether these two phenomena can cause the problems of convergence and unacceptable solutions reported above for the Russian data? This point will be discussed on the basis of a simple

Pearson correlation matrix obtained for the nationality questions. The same question is asked four times. In the two successive interviews, leading to a four by four correlation matrix shown in table 5.

TABLE 5: Pearson correlation coefficients between repeated measurements of the second trait (*Anyone who lives and works in Russia has the right to the Russian nationality.*) of the nationality topic.

	R_{231}	R_{232}	R_{241}	R_{242}
R_{231}	1.00			
R_{232}	.81	1.00		
R_{241}	.14	.05	1.00	
R_{242}	.14	.15	.81	1.00

If this correlation matrix is analysed with a simple model, consisting of two correlated trait factors (one for each interview) each with two indicators (from the same interview), and assuming uncorrelated error terms, it can easily be verified that the estimation procedure does not converge.

If the argument about crystallisation of an opinion is correct, then either the correlation within the interviews is too high, or the correlation across the interviews is too low, or both. This point can be verified on this example by simply changing the correlation found in the indicated directions and testing to see whether the convergence problem and the unacceptable solutions then disappear.

When this was done, we found that a reduction of the correlation coefficients within the interviews from the .81 actually found to .61, instead of .81 as it was found, is sufficient to obviate the problems of non-convergence and improper solutions. On the other hand, it was also found that these problems disappear if the across wave correlation coefficients are increased by .1. These results are in accord with our belief that the problems are due to a combination of events in these cases, viz. the relatively high correlation within the interviews and the relatively low correlation across the interviews. If only one or neither occur, no problems arise. This raises the question of why the correlation is either too low or too high, or both. This will be dealt with in the next section.

4. POSSIBLE EXPLANATIONS

The first question we want to discuss is why the correlation across the interviews is so low in the Russian data (compared with the Dutch data). The answer to this question would appear to be rather simple at first glance. The time between the interviews in Russia is approximately a year, which is much longer than in the Netherlands where there is often only two weeks between the interviews. This might be sufficient to explain why the across wave correlation coefficients differ so much between the two countries, at least where people have no crystallised opinion on the issue. Where opinion is not crystallised the stability (correlation) could be very low, as was also found by Converse (1964). On the other hand, if opinion is not crystallised one would not expect such high correlation within the interviews. Hence, the idea that the low correlation found across the interviews is due to uncrystallised opinion contradicts with the very high correlation found within the interviews, at least for topics were we expect less awareness of the respondents. This argument - that the low across wave correlations are due to uncrystallised opinion - can only be correct if there are good reasons for expecting the correlation between the same variables within the interviews to be too high. From this it follows that the correlation between repeated observations of the same variable within an interview is caused not only by the trait of interest, but by some other factor(s) too.

The second question is why the correlation within the interviews is so high in the Russian data (compared with the Dutch data). Our conclusion on the first question suggests that there are other factors at work that cause extra correlation between repeated observations of the same variable within an interview in Russia. But if this is so, why not in The Netherlands? We do not say that these factors are not at work in the Dutch data, merely that they are not as fatal to the Dutch data as they are to the Russian data. This is probably due to specific factors in the design of the multitrait multimethod experiments. The question is then which factors in a multitrait multimethod design can cause extra correlation between repeated observations of the same variable within an interview, especially in the design implemented in Russia.

There are several design factors that could cause extra correlation between repeated observations. The first is the similarity of the measurement procedures. It is assumed that the covariance between the error terms of repeated observations of the same variable is zero. This assumption is only justified if the measurement procedures are dissimilar, regardless of what exactly one understands as dissimilar measurement procedures (De Wit and Billiet, 1995). The second factor is acquiescence. Acquiescence - the inclination to agree on approval statements regardless of the content - cannot be avoided if approval statements are used (Krosnick and Fabrigar, forthcoming). In that case, extra correlation will arise when observations of the same variable in one interview are repeated with approval statements. The third factor is the presence or absence of memory effects. It is assumed there will be no memory effects if the time between the repeated observations is at least twenty minutes, and if the questions between the repetitions are of the same kind, and if the response given to the first question is not extreme (Van Meurs and Saris, 1995). All three factors mentioned above will be evaluated as far as possible on the basis of the available Russian data, thus leaving the Dutch data for what they are.

4.1. Evidence that correlation is inflated by the use of similar measurement procedures

To find evidence that similar measurement procedures do indeed cause extra correlation between repeated measures of the same trait, we looked for topics in the data set that were measured with similar measurement procedures within one interview and with dissimilar measurement procedures in another. Some topics, where one of these forms was not available, where included for they were of sufficient interest to study them in the light of the awareness argument presented above. The results are shown in table 6. Again we distinguish between topics with respect to people's awareness of the topic. We assume that the topics are ordered according to degree of awareness from high (top) to low (bottom), though this arrangement should not be taken too strictly. What is important to note is merely that respondents are more aware of topics in the upper part of the table than in the lower part.

It is easy to verify from table 6 that the correlation between repeated observations of the same variables is higher if they are observed using similar measurement procedures. Moreover, the correlation remains stable across topics, whereas the correlation between variables observed with dissimilar measurement procedures show a pattern of decreasing correlation coefficients, which accords with the notion that a decrease in awareness is associated with a decrease of correlation (opinion). On the basis of this information we can only conclude that similar measurement procedures inflate correlation between repeated observations of the same trait within an interview.

In addition, we believe that the distortion is more severe when the degree of awareness decreases. This can be verified in table 6 by comparing the correlation in the columns showing dissimilar and similar measurement procedures for each topic. For topics where we would expect respondents to be more aware of the issue at stake the differences in the correlation coefficients are rather small, except for the topic 'buying' [i]. For example, the difference for 'satisfaction' between similar and dissimilar measurement procedures in the lowest correlation is 0.08, and in the highest correlation the difference is 0.10. The differences for 'political efficacy' are already much larger. The difference for 'political efficacy' in the lowest correlation is .24 and in the highest correlation the difference is .21, which in both cases twice as large as the differences for 'satisfaction'. This is why we believe that the severity of the distortion increases as the degree of awareness decreases.

TABLE 6: Pearson correlation coefficients between the same variables within an interview. A star in a cell indicates that no information is available[ii].

[i] The correlation coefficients are produced by a forced choice question with four scale points and an approval statement with a 101 point rating scale, which is extremely different in this context.

[ii] The following rule has been applied to decide whether measurement procedures are similar or dissimilar. If measurement procedures do not differ too much with respect to the number of scale points

Correlation between the same	Dissimilar		Similar[i]	
	Low	High	Low	High[ii]
Buying	.10	.29	.79	.89
Satisfaction	.65	.74	.73	.84
Change	.58	.78	*	*
Political efficacy	.34	.60	.58	.81
Ingroup	.31	.57	*	*
Political interest	.12	.53	.81	.87
Nationality	*	*	.67	.81

In sum, the correlation between repeated measurements of the same trait will be inflated if similar measurement procedures are used, and even more inflated if the awareness is lower.

4.2. Evidence for the presence of acquiescence

Acquiescence is defined as the endorsement of an assertion made in a question regardless of the content of that assertion. It presumably becomes manifest as an inclination to say 'agree' when people are given an agree/disagree (or true/false, or yes/no) set of response choices. Acquiescence distorts both the observed distributions of responses and the correlation between variables. It can be responsible for a considerable proportion of shared variance across different variables (Krosnick and Fabrigar, forthcoming). If that is true, then it will also distort the correlation between the repeated observations of the same variable (in particular within an interview), given that both are measured with an approval statement and with (very) similar response scales. This is true for most topics for which we assume that respondents have little awareness. Table 7 presents for several topics the percentage of 'completely agree' responses, the most extreme answer a respondent could give. All scales provided at least five scale points. The scale for 'buying' was a 101-point scale; values from 0 till 20 were counted as 'completely agree'.

TABLE 7: Percentage of ‘completely agree’ responses for several topics for which approval statements were employed.

and the questions are either both in a forced choice form or both in the form of an approval statement, then we called the measurement procedures similar. In any other case we called them dissimilar.

[i] Similarity and Dissimilarity refers to the similarity/dissimilarity of the scales that are used in the measurement procedure for each topic.

[ii] Low and High refer to the lowest and highest correlation computed for each topic=set of variables.

	Variable					
Percentage Completely agree for	R1	R2	R3	R4	R5	R6[i]
Buying	43.7	30.4	30.8	36.6	35.0	34.5
Political efficacy	58.5	40.0	41.9			
Nationality	61.6	19.1	55.2	48.6	44.8	
Ingroup	53.2	43.4	29.8	41.3		

The figures in the table clearly show that the percentage of 'completely agree' responses appears excessive, without taking into account the moderate 'agree' responses. This seems to be a solid indication of acquiescence in these data. One may, however, think that these percentages actually reflect real opinion. This can be tested for 'buying', where we also asked the respondents in the same interview with a forced choice question whether they bought goods at a specific selling point 'often', 'sometimes', 'rarely' or 'never'. The forced choice answer 'often' corresponds to 'completely agree' in the approval statement. We summarised the percentages of the extremes, 'often' and 'never', and the percentage 'completely agree' in table 8. The number of cases was approximately 2200 for each variable, and varied only due to partial non-response.

TABLE 8: Forced choice response extremes compared with 'completely agree' responses for buying.

Buying	% Never	% Often	% Completely Agree
R1	2.3	76.9	43.7
R2	27.1	13.8	30.4
R3	41.8	10.4	30.8
R4	14.1	35.8	36.6
R5	51.5	7.1	35.0
R6	97.3	.3	34.5

For the observed variable R6, the category 'never' (buy goods at that specific selling point) attracts almost all respondents, but when the same respondents have to answer the approval statement almost 35% of the respondents agree that they 'often' buy goods at that specific selling point. This is a remarkable difference. Also for the observed vari-

[i] For each topic several questions on different aspects were asked in the form of a battery. Thus Ri refers to the Rth observed variable (question) of a topic.

ables R2, R3, and R5 a similarly remarkable change of judgement has taken place in only half an hour. We find this convincing evidence for acquiescence in these data.

To summarise, acquiescence inflates the correlation between repeated observations of the same variable when approval statements and similar response scales are used in both instances. When it comes to the respondents' awareness, one could have reservations about the comparability of 'buying' on the one hand and 'political efficacy', 'nationality', and 'ingroup' on the other. But acquiescence is the endorsement of an assertion made in a question, regardless of the content of that assertion, so that such reservations in no way affect the question under consideration, viz. the presence of acquiescence.

4.3. Evidence that memory effects are present

Memory can have effects in two different ways. The simplest memory effect is that respondents recall the response they gave to the first question. This is rather unlikely in the Russian interviews. For example, in the third interview where we repeated about 40 questions, it would take an extraordinary good memory to remember all or even some of the previous responses. This recall may be unlikely for most responses, however, but not necessarily for all. Van Meurs and Saris (1995) have shown that respondents who give an extreme answer on the first measurement are indeed capable of recalling their previous answer in the same interview, regardless of the time between the observations. This response consistency is even more evident with those people who are very involved in the issue and have therefore a crystallised opinion. In such a case, it is however unclear whether one should speak of a memory effect, since the response reflects a crystallised opinion. On the other hand, if people are not involved in the issue and thus probably do not hold a crystallised opinion, it is clear that one should call it a memory effect. In the previous section on the effect of acquiescence, we showed that a considerable percentage of respondents chooses the extreme category, i.e. 'completely agree'. The low response consistency that was found across the interviews for those people who choose the extreme category indicates that they do not have a crystallised opinion. It is therefore very likely that the high consistency within the interview is due to the recalling of their previous response rather than an independent expression of a stable opinion.

There is a second reason why we should expect memory effects. A memory effect may also be a reproduction of the response process (Sudman et al, 1996; Tourangeau and Rasinski, 1998), or rather the sequence of processes by which respondents generate an answer. These processes are generally question-interpretation, information retrieval, information integration, and finally converting the judgement to a response alternative on the presented scale. If a respondent's opinion on a particular topic is not crystallised, then the response process will hinge mostly upon volatile information that is immediately salient to the respondent. As a consequence, the respondent will probably make little effort to think carefully about the question, and moreover he or she is probably not capable of doing so because he or she does not have a structured pile of information to form a considered opinion. Hence, the response process will be very simple. If that respondent is again confronted with the same question, the same context, and a similar response scale in one interview, he or she may well remember the way the answer was produced on the first occasion, triggering the same simple response process and in consequence a response that is very close to the previous response, if not identical. We have no direct evidence that this kind of memory effect occurs but it is not unlikely, particularly in the light of our findings with similar measurement procedures. Again, as in the case of acquiescence, similarity of measurement procedures seems to be a condition for this type of memory effect to occur. In addition, under this special condition the notion of a memory effect can be extended to a more general notion of memory effects that is not restricted to respondents with a low degree of crystallisation of an opinion. Such a type of memory effect will be triggered for any kind of variable when in the repeated observation the context, the formulation, and the response scale of the question are highly similar.

In summary, the first type of memory effect, direct recalling of the answer, is likely to play a role with respondents who give extreme answers. Evidence for this was presented in the section on acquiescence. The second type of memory effect, reproducing the response process, is likely to occur when in both observations of the same variable within an interview, the context and response scale are very similar. Evidence for such type of memory effect was presented in the section dealing with similar measurement procedures. The correlation between repeated observations of the same variable will be inflated due to these memory effects when similar measurement procedures are used. In addition, the lower the respondents' degree of awareness of a particular topic, the more the correlation will be inflated, and vice versa.

In conclusion, four factors are found to cause extra correlation between repeated observations of the same variable within an interview: similarity of measurement procedures, acquiescence, recalling of extreme answers, and reproduction of a response process. Both acquiescence and reproduction of the response process are triggered when in both observations the measurement procedures are very similar.

5. CONCLUSION

In an attempt to estimate the quality of survey measurement instruments used in the Russian multitrait multimethod study, the data were analysed using the true score multitrait multimethod model and the correlated uniqueness model. It appeared that both these models are unsuitable for these data. Although for some topics the analyses yielded acceptable results, for most topics they did not. We encountered two main problems, non-convergence of the model, and unacceptable estimates (negative variances), particularly with topics where we assumed that respondents had a low degree of awareness of the issue at stake.

Since the same design was used in The Netherlands without leading to so much problems, we examined the correlation between variables that were comparable with respect to degree of awareness of the topic. Comparison of these correlations led to the conclusion that something might be wrong with the correlation between the same variables within the interviews in Russia. That is, we found that the correlation between the same variables within the interviews is - for low awareness topics - excessive compared with the correlation across the interviews. Suggesting on the one hand stable (crystallised) opinions within the interview and on the other hand (very) unstable opinions across the interviews.

Having concluded that the correlation between the same variables within interviews was too high in Russia, explanations were sought for the origin of problems encountered in the data analyses. Factors were sought in the multitrait multimethod design that could cause violations of the assumptions of the models used to analyse the data. Four factors were found which, if not met, could produce extra correlation between repeated observations of the same variables within the interviews. The first factor was the (dis)similarity of the measurement procedures. If measurement procedures are too similar, the assumption for the true score model that the method factors are not correlated will be violated. Or, in the case of the correlated uniqueness model, the assumption that the covariance between the error terms of observations of the same variable is zero will be violated. The second factor is acquiescence. When similar measurement procedures are used in both observations of the same variable in one interview, this factor will play a role. The third factor is the reproduction of the response process (memory effect). Again, this factor will play a role when similar measurement procedures are used. The fourth factor is the recall of extreme answers (memory effect). If memory effects are present, the assumption of zero covariance between the error terms of observations of the same variable is violated in both the true score and the correlated uniqueness model.

We looked to see whether the design of the Russian multitrait multimethod study provided a context so that these assumptions are violated. It followed that the correlation between the same variables within the interviews were higher when (very) similar measurement procedures were used. We also found that when approval statements were used to measure opinions, acquiescence was present. Furthermore we expected memory effects to occur in two instances. Firstly, if the measurement procedures are very much the same for repeated observations then respondents will recognise the context of the questions which makes it rather easy to reproduce their previous answer. Secondly, Van

Meurs and Saris (1995) have shown that respondents who give extreme answers are very capable of recalling their previous answer.

We are now convinced that a combination of memory effects, acquiescence, and similar methods distort the correlation, i.e. cause extra correlation, between repeated observations of the same variable within an interview. Inspection of the data confirms the effects of acquiescence and similar methods. The evidence for memory effects may be circumstantial. However, all these distorting factors are interrelated and have the same effect on the correlation between repeated observations of the same variable within an interview. It is therefore difficult to formulate a test that can differentiate between these effects (factors).

Are these distorting factors the reason why the multitrait multimethod analyses failed in many cases? The answer is 'not entirely'. The failure is due to a combination of the factors mentioned above, and by unstable opinions, i.e. the low correlation across the interviews. This follows from the fact that the multitrait multimethod analyses hardly produce any problems for topics where respondents have well crystallised opinions, despite the fact that correlated errors, due to the use of similar measurement procedures, acquiescence, and memory effects, might and probably do occur. Problems in the data analyses only surface when people have no sufficiently crystallised opinion on an issue to form a stable opinion. In that case the correlation across the interviews will obviously be very low, whereas the correlation within the interviews artificially inflates, as a result of the factors mentioned. It is this specific combination in the Russian surveys that leads to the problems in the multitrait multimethod analyses.

All in all, it is evident that the correlation across the interviews is not artificially low but the correlation within the interviews is artificially high because of the factors discussed above. If this is a correct analysis of the problem, one should correct for these distorting factors by introducing correlated error terms between repeated observations of the same variable within the interviews in the multitrait multimethod models. Unfortunately, if such correlated error terms are introduced into these models, the models are not identified. This is therefore not a feasible solution. What is required is the possibility of estimating the magnitude of these distorting factors (correlated error terms) in another context. In addition one needs to look for groups in the population with respect to the degree of awareness about specific topics. This way, it can be tested whether such groups differ in both the magnitude of the model parameters and the factors discussed above. This presents an interesting challenge for further research.

With respect to surveys held in Russia one should be aware of these problems. Especially lack of opinion and overestimation of reliability and acquiescence. In cross-sectional surveys these problems can lead to highly overestimated correlations between substantive variables. It is not unlikely that these problems are also present in Western surveys, but in a less extreme form, nevertheless, it means that also these surveys require correction for these problems.

REFERENCES

Anderson, J. C., and D. W. Gerbing. 1984. 'The effect of sampling error on convergence, improper solutions and goodness-of-fit indices for maximum likelihood confirmatory factor analysis.' *Psychometrika* 49: 155-173.

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

Billiet, J., G. Loosveldt, and L. Waterplas. 1986. 'Het survey interview onderzocht. Effecten van het ontwerp en gebruik van vragenlijsten op de kwaliteit van de antwoorden.' *K.U.L.*. Leuven.

Browne, M.W. 1984. 'The decomposition of multitrait multimethod matrices.' *Brittish Journal of Mathematical and Statistical Psychology* 37: 1-21.

Campbell, D. T., and D. W. Fiske. 1959.'Convergent and discriminant validation by the multimethod multitrait matrix.' *Psychological Bulletin* 56: 833-853.

Coenders G., and W. E. Saris. Forthcoming. 'Testing nested additive, multiplicative and general multitrait-multimethod models.'

Converse, P. E. 1964. 'The nature of belief systems in mass opinion.' Pp. 206-261 in *Ideology and Discontent*, edited by D. E. Apter. London: Collier-MacMillan Ltd.

Heise, D. R., and G. W. Bohrnstedt. 1970. 'Validity, invalidity, and reliability.' In *Sociological Methodology*, edited by E.F. Borgatta, and G.W. Bohrnstedt. San Francisco: Jossey-Bass.

Jöreskog, K.G., and D. Sörbom. 1993. 'LISREL8: User's Reference Guide.' Chicago (Il.): Scientific Software International.

Jöreskog, K.G., and D. Sörbom. 1996. 'PRELIS2: User's Reference Guide.' Chicago (Il.): Scientific Software International.

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

Költringer, R. 1995. 'Measurement quality in Austrian personal interview surveys.' Pp. 207-224 in *The Multitrait Multimethod approach to Evaluate Measurement Instruments*, edited by W. E. Saris and A. Münnich. Budapest: Eötvös University Press.

Krosnick, J. A., and L. R. Fabrigar. Forthcoming. *Designing Good Questionnaires: Insights from Cognitive and Social Psychology*.(tentative title) New York: Oxford University Press.

Lodge, M., M. R. Steenbergen, and S. Brau. 1995. 'The responsive voter: campaign information and the dynamics of candidate evaluation.' *American Political Science Review* 89: 309-326.

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

Luijben, T. 1989. *Statistical guidance for model modification in covariance structure analysis*. Amsterdam: Sociometric Research Foundation.

Luskin, R. C. 1997. 'Political psychology, political behaviour, and politics: questions of aggregation, causal distance, and taste.' In *Thinking about Political Psychology*, edited by J. H. Kuklinski. New York: Cambridge University Press (forthcoming).

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

Marsh, H. W., and M. Bailey. 1991. 'Confirmatory factor analysis of multitrait multimethod data: comparison of the behaviour of alternative models.' *Applied Psychological Measurement* 15: 47-70.

Meurs A. Van, and W. E. Saris. 1995. 'Memory effects in MTMM studies.' Pp. 89-102 in *The Multitrait Multimethod approach to Evaluate Measurement Instruments*, edited by W. E. Saris and A. Münnich. Budapest: Eötvös University Press.

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

Saris, W. E. 1982. 'Different questions, different variables.' Pp 78-96 in *A Second Generation of Multivariate Analysis*, Vol. 2. *Measurement and Evaluation*, edited by C. Fornell. New York: Preager.

Saris W. E., and F. M. Andrews. 1991. 'Evaluation of measurement instruments using a structural modelling approach.' Pp. 575-598 In *Measurement Errors in Surveys*, edited by P. P. Biemer, R. M. Groves, L. E. Lyberg, et al. New York: John Wiley.

Saris W. E., and J. van der Zouwen. 1999. 'Feasibility study concerning a Ressearch Program on the quality of Survey Measurement.' N.W.O. Department of MAGW (Dutch organisation for Scientific Research).

Saris W. E., and W.M. van der Veld. 2000. 'A Program for prediction of the Quality of Survey Measurement'. Paper to be presented in October 2000 at the methodology conference in Köln, Germany.

Scherpenzeel A. C., and W. E. Saris. 1995 'The quality of indicators of statisfaction across Europe: a meta analysis of multitrait multimethod studies.' Pp. 51-73 In *A Question of Quality: Evaluating Survey Questions by Multitrait-Multimethod Studies.* Edited by Scherpenzeel A. C. Amsterdam, The Netherlands: University of Amsterdam, Statistics and Methodology Department.

Scherpenzeel A. C., and W. E. Saris. 1997. 'The validity and reliability of survey questions: a meta-analysis of multitrait multimethod studies.' *Sociological Methods & Research* 25: 341-383.

Schuman, H., and S. Presser. 1981. 'Questions and answers in attitude surveys: Experiments on question form, wording and context. New York: Academic Press.

Sniderman, P. M., R. A. Brody, and P. E. Tetlock. 1991. *Reasoning and Choice: Explorations in Political Psychology*. New York: Cambridge University Press.

Sudman, S., N. M. Bradburn, and N. Schwarz. 1996. *Thinking about answers, the application of cognitive processes to survey methodology*. San Francisco: Jossey-Bass.

Tourangeau, R., and K. A. Rasinski. 1988. 'Cognitive processes underlying context effects in attitude measurement.' *Psychological Bulletin* 103: 299-314.

Wit, H. de, and J. Billiet. 1995. 'The MTMM design: back to the founding fathers.' Pp. 39-59 in *The Multitrait Multimethod approach to Evaluate Measurement Instruments*, edited by W. E. Saris and A. Münnich. Budapest: Eötvös University Press.

Zaller, J. R. 1992. *The nature and origins of mass opinion.* Cambridge: University Press.

Ontvangen: 7 januari 2000
Geaccepteerd: 11 mei 2000

Appendix 1: The data preparation phase.

All models were fitted to Pearson correlation matrices. The correlation coefficients were estimated with PRELIS2 (Jöreskog and Sörbom, 1996). Table 1A shows the method used to cope with missing values for each topic, and also the sample size specified for each topic. In the case of pairwise deletion, an average of the number of cases across the different correlation coefficients in the matrix has been used. For some topics the number of cases seems to drop dramatically. This is not due to item non-response, although item non-response was occasionally found it is not a serious cause of item non-response. There are two main reasons for (item) non-response. The first is that cases are missing by design, that is the multitrait multimethod experiments with 600 cases or less have been conducted within randomly selected subgroups (N = approx. 700) of the whole sample. Furthermore, only native Russians were allowed to participate in these multitrait multimethod experiments. Secondly, non-response has been caused by panel attrition (table 2A). Note, that for most topics were listwise deletion is applied the N drops dramatically, that is because we computed the correlations for all available waves (not just those with the experiments). Finally, for the topic spending serious data cleaning has been done, which explains the low sample size used.

Table 1A: The sample size used for each topic, and the deletion method used.

Topic	Start size of sample	Sample size used	Deletion method
Political efficacy	3728	2277	Listwise
Change	2807	1504	Listwise
Satisfaction	2807	1257	Listwise
Spending	2807	802	Listwise
Buying	2272	1465	Listwise
Ingroup	2272	500	Pairwise
Nationality	2272	600	Pairwise
Outgroup	2272	500	Pairwise
Policy	2272	500	Pairwise
Threat	2272	500	Pairwise
Trust	2272	600	Pairwise

Table 2A: The number of respondents for which a completed interview has been obtained.

Wave	1	2	3	4
Completed interviews	3728	2807	2272	2074