UNIVERSITAT POMPEU FABRA

RECSM

Research and Expertise Centre for Survey Methodology

RECSM Working Paper Number 19

2011

Universitat Pompeu Fabra - Research and Expertise Centre for Survey Methodology
Edifici de França - Despatxs 70.374-70.382
Passeig de Circumval·lació, 8 - 08003 Barcelona
Tel.: +34 93 542 11 63 web: http://www.upf.edu/survey/ e-mail: recsm@upf.edu

The split-ballot multitrait-multimethod approach:

Implementation and problems

Melanie Revilla

Willem Saris

Universitat Pompeu Fabra, RECSM

Abstract:

Saris, Satorra and Coenders (2004) proposed a new approach to estimate the quality of survey questions, combining the advantages of two existing approaches: the multitrait-multimethod (MTMM) and the split-ballot (SB) ones. Implemented in practice, this new approach led to frequent problems of non-convergence and improper solutions. This paper uses Monte Carlo simulations to understand why the SB-MTMM is working well in some cases but not in others. The number of SB groups is a crucial element: the 3-group design is performing better. However, the 2-group design can also perform well: the analyses suggest that the interaction between the absolute values of the correlations between the traits and the relative values of the different correlations between traits play an important role.

Key words:

Split-ballot multitrait-multimethod approach, convergence, Heywood cases, Monte Carlo simulations, quality of survey questions

1. Introduction

The idea of repeating questions using different methods started with Campbell and Fiske (1959, p. 81): "In order to examine discriminant validity, and in order to estimate the relative contributions of trait and method variance, *more than one trait* as well as *more than one method* must be employed in the validation process […] it will be convenient to achieve this through a multitrait-multimethod matrix. Such a matrix presents all of the intercorrelations resulting when each of several traits is measured by each of several methods." They suggested a direct analysis of all these intercorrelations.

At the beginning of the 1970's, the MTMM approach saw new developments: Werts and Linn (1970) and Jöreskog (1970, 1971) proposed to treat the MTMM matrix as a Confirmatory Factor Analysis model. In the 1980's, Andrews (1984) suggested using the MTMM approach to evaluate the reliability and validity of single questions. Browne (1984) and Cudeck (1988) proposed a multiplicative method effect model, whereas Saris and Andrews (1991) suggested to use a true score model. During the 1990's and 2000's, many authors continued using and discussing this approach (e.g. Költringer, 1993; Scherpenzeel, 1995; Wothke, 1996; Alwin, 1997 and 2007; Eid, 2000; Saris and Gallhofer, 2007; Revilla, Saris, Krosnick, 2009).

Different papers (Corten et al., 2002; Saris and Aalberts, 2003) showed that the true score model performs better than the alternative models. For this reason, it is the model used in Saris et al. (2004) and also in this paper. For identification, at least three traits need to be repeated using at least three methods. When considering three traits and methods, the true score model can be defined by the following system of equations (Saris and Andrews, 1991):

$$Y_{ij} = r_{ij} T_{ij} + e_{ij} \quad \text{for } i = 1,2,3 \text{ and } j = 1,2,3 \qquad (1)$$

$$T_{ij} = v_{ij} F_i + m_{ij} M_j \quad \text{for } i = 1,2,3 \text{ and } j = 1,2,3 \qquad (2)$$

Where: Y_{ij} is the observed variable for the i^{th} trait and the j^{th} method; r_{ij} and v_{ij} are respectively the reliability and validity coefficients for the i^{th} trait and the j^{th} method; T_{ij} is the systematic component or true score of the response Y_{ij}; e_{ij} is the random error associated with Y_{ij}; F_i is the i^{th} trait (or factor); M_j is the variation in scores due to the j^{th} method; m_{ij} is the method effect for the i^{th} trait and the j^{th} method.

It is assumed that: the traits are correlated with each other, the random errors are *not* correlated with each other, nor with the independent variables in the different equations, and the method factors are *not* correlated with each other, nor with the traits or the random errors. The method effects for one specific method M_{j*} are equal for the different traits. In practice, the m_{ij} are set equal to one, whereas the method variances are left free. In contrast, the traits' variances are set to one, whereas their loadings are left free.

A graphical representation of this true score model is presented in Figure 1. For clarity only the first observed variable Y_{11} is represented, but each true score T_{ij} should similarly be linked to an observed variable Y_{ij} through the reliability coefficient $r_{ij.}$

Figure 1: The true score model for 3 traits and 3 methods

When the model of Figure 1 is completely standardised, $r_{11,}$ v_{11} and m_{11} correspond respectively to the reliability, the validity and the method effect coefficients for the first trait measured with the first method.

Estimates for all parameters of the model can be obtained using classical Structural Equation Modelling softwares. The total quality of a measure is then computed as: $q^2_{ij} = r^2_{ij} * v^2_{ij.}$ It represents the strength of the relationship between the observed variable Y_{ij} and the latent variable of interest F_i or said differently it represents the explained variance in the observed variable by the latent trait of interest. The quality is in the [0,1] interval: the closer to one, the better the measurement instrument.

The split-ballot MTMM approach (SB-MTMM)

Saris et al. (2004) argue that repeating three times the same question increases the cognitive burden of the respondents and threatens the accuracy of the measurements. Since at least 20 minutes of similar questions (Van Meurs and Saris, 1990) should separate one question from its repetition in order to avoid memory effects, quite long interviews are also needed to implement the MTMM approach. Because of these limitations, Saris et al. (2004) propose to combine MTMM and split-ballot (SB) designs.

In the SB-MTMM approach, different groups are drawn randomly from the population. Each group gets only two methods, but different groups get different combinations of methods. In that way the number of repetitions for each respondent is reduced but information is obtained about the three methods. In Saris et al.´s words (2004, p.313): "It enables researchers to evaluate measurement reliability and validity, and does so while reducing the response burden." They suggest that the estimates can be obtained by Maximum Likelihood (ML) estimation for multiple-group (MG) analyses.

Different SB-MTMM designs can be thought of. We focus on the 2-group and 3-group designs that are the most common in practice. A possible 2-group design is: group 1 gets method 1 (M_1) at time 1 and method 2 (M_2) at time 2, whereas group 2 gets method 1 (M_1) at time 1 but method 3 (M_3) at time 2. Different assignments of the methods to the groups are possible: for instance in group 2 respondents could have M_2 at time 1 instead of M_1. For simplicity, we stick to the 2-group design just mentioned.

The specificity of the 2-group design is that the combination of methods 2 and 3 is missing by design. This means that the set of correlations between the variables measured with M_2 and the variables measured with M_3 is completely absent.

Saris et al. (2004) show that in the 2-group design the model is identified "under normal circumstances" (p. 332) and that the "estimation procedure specified [ML for MG] will provide consistent estimates of the population parameters." The authors identify three cases where empirical identification may not be obtained: "minimal variance of one of the method factors", "lack of correlation between the latent traits" and "equal correlations between the latent traits" (p. 333). Excluding this, they argue that the 2-group design performs well (meaning converges and provides accurate estimates) even if larger sample sizes may be needed to get the same accuracy as with the 3-group design.

A possible 3-group design is: group 1 gets M_1 at time 1 and M_2 at time 2, group 2 gets M_2 at time 1 and M_3 at time 2, and group 3 gets M_3 at time 1 and M_1 at time 2. Here the correlations between M_2 and M_3 are present in group 2, so information is available for all blocks of the MTMM matrix. According to Saris et al. (2004), no empirical identification problems are expected.

In the next section, examples of real data analyses are used to illustrate the different kinds of problems encountered when analysing real data of SB-MTMM experiments and their frequency. Because of these difficulties in real data analyses, section 3 presents Monte Carlo simulations used to investigate under which conditions the SB-MTMM performs well and under which conditions it does not. Section 4 draws conclusions and provides possible lines for further research.

2. Implementation of the SB-MTMM approach

The 2-group design has been implemented on a large scale by the European Social Survey[1] (ESS) since 2001. The ESS uses in each round a 2-group SB design to collect

[1] For more information, please see: http://www.europeansocialsurvey.org/

data for several MTMM experiments in around 30 countries. The survey is divided into a main questionnaire (same for all respondents: M_1), and a supplementary questionnaire, that differs for the two SB groups (M_2 in group 1, M_3 in group 2, cf. section 1).

The 3-group design has also been implemented: for instance, in December 2008, the Longitudinal Internet Studies for the Social Sciences (LISS[2]) panel presented to its respondents a survey including some 3-group SB-MTMM experiments. The 3-group design however is more difficult to implement. Indeed, in the 2-group design the methods differ only at time 2 for the different SB groups, whereas in the 3-group design, they differ both at times 1 and 2. It is not possible with the 3-group design to have one main questionnaire similar for all respondents. Preparing the survey is therefore more demanding. Besides, researchers who want to analyse identical questions can only use two out of the three SB groups, so it reduces their sample size. Even if it concerns only the variables included in the MTMM experiments, many survey institutes prefer to use the 2-group designs. However, this leads to recurrent problems in the analyses.

2.1. Main problems encountered in practice

Rindskopf already remarked in 1984 that in practice "structural equation models are often plagued by a variety of undesirable results" (p.109). He argued it was a consequence of empirical underidentification: "for most models, one cannot say that the model is identified but only that it may be identified if certain conditions are true. [...] The conditions for identification generally take the form of requiring that certain parameters not to be zero or that parameters not equal one." (p. 110). If these conditions are not satisfied in a specific dataset, then undesirable results may arise.

[2] Dutch Web panel based on probability sample. For more information, please see: http://www.centerdata.nl/en/LISSpanel

Since 1984, much work has been done on structural equation models, but the issue of undesirable results is still present. For the SB-MTMM model, the "undesirable results" take mainly two forms: non-convergence (NC) and Heywood cases (HC). HC or "improper solutions" correspond to "negative variances or correlation estimates greater than one in absolute value" (Kolenikov and Bollen, 2008, p.1). Biased estimates may also be an issue but without knowing the true values, it is difficult to notice it.

NC is problematic since if the parameters cannot be estimated, no conclusion can be drawn. HC are also problematic. Negative variances may appear just because of sampling fluctuations if the true value of the parameter is close to zero (Van Driel, 1978). That is why it is often argued that HC can be simply solved by fixing to zero the negative but non-significant values. However, Rindskopf (1984) underlined that "the corrective action to take is not always obvious; for example, it is not always correct to remove a parameter from an analysis when it has negative error variance estimate, because the problem may be caused by another variable" (p. 110).

Despite this warning, fixing the negative non significant values to solve HC is a quite common procedure, implemented for example in Saris et al. (2004, p. 331).

Nevertheless, our analyses of real SB-MTMM data are in line with Rindskopf's comment and suggest that fixing negative non significant estimates may have a large impact on other estimates of the model and may not really be a solution. This can be illustrated by the 2-group SB-MTMM experiment about satisfaction in the Netherlands collected in the first ESS round (2002-2003). The three traits deal with satisfaction with the "present state of the economy", the "way the government is doing its job" and the "way democracy works in the country". In the main questionnaire, respondents get an 11-point scale going from "extremely dissatisfied" to "extremely satisfied" (M_1). In the supplementary one, group 1 gets a 4-point scale going from "very dissatisfied" to "very

satisfied" (M_2), whereas group 2 gets a 6-point scale going from "extremely dissatisfied" to "extremely satisfied" (M_3).

The covariance matrices are analysed using ML estimation for MG in LISREL[3] (Jöreskog and Sörbom, 1991). The model used is the true score model (cf. introduction). Since the respondents are randomly assigned to the SB groups, one does not expect significant differences across groups for the same questions, so the parameters in the second group are specified invariant (details in Saris and Gallhofer, 2007, chapter 12).

The estimation of this satisfaction experiment in the Netherlands leads to a HC: the method variances for M_2 and M_3 are negative, but according to a t-test not significant. We start by fixing M_2 variance. The variance of M_3 being still negative, we also fix it and get a proper solution (PS).

To determine if the model appropriately reproduces our data, we use the software JRule (Van der Veld et al., 2008) based on the testing procedure developed by Saris et al. (2009). Using information about types I and II errors, it provides a test for misspecifications at the parameter level. According to JRule, the method variances fixed are not misspecified and the model cannot be rejected. This seems to provide support to the procedure of fixing negative variances. However, instead of M_2 and M_3 variances, we could also fix the variance of M_1. This is an alternative way of getting a PS. Also this model cannot be rejected. In particular, no misspecifications are found for the method variance fixed (variance of M_1).

Even if we got proper solutions in both cases, the results seem determined by the choice of fixing one or the others method variances. The 11-point scale for instance (M_1) has the lowest quality of all three methods in the first situation (when fixing φ_{55} and φ_{66}), but the highest in the second one (when fixing φ_{44}).

[3] An example of Lisrel input to analyze SB-MTMM experiments is available online: http://bit.ly/gQI3sV

One could argue that the first model is the good one: fixing a non significant parameter seems more acceptable than fixing the only positive and significant variance. However, the second model cannot be rejected according to JRule[4]. We are more willing to think that getting so different estimates with two fitting models suggests that both sets of estimates are biased because of the decision of fixing some parameters. So getting HC may really lead to problematic situations where it is not clear what to do.

2.2. Frequency of these problems

NC and HC are all the more problematic as they are occurring very frequently. The first and fourth rounds of the ESS are used to illustrate this. In the first round, six SB-MTMM experiments (with three traits and three methods) dealing with media use, political efficacy, political orientation, satisfaction, social and political trust are analysed in 19 countries. In the fourth round, three SB-MTMM experiments dealing with media use, satisfaction and political trust are considered. 22 analyses are run based on the country and language of the interview. In total, 6*19 + 3*22 = 180 SB-MTMM experiments are therefore studied[5].

Table 1 reports the number of NC, HC and PS. One can notice that for the NC cases, one does not know if solving the non-convergence would lead or not to a proper solution.

Table 1: Results obtained when running 180 SB-MTMM models for ESS rounds 1 and 4

Table 1 shows that only in 23.3% of the datasets a proper solution is obtained, whereas 30.0% of the datasets lead to non convergence and 46.7% to Heywood cases.

[4] The two models are also very similar and cannot be rejected if we consider more global tests of the model as the Chi-square or fit indices as RMSEA.
[5] For more details about the traits and methods used in each experiment, as well as for the list of countries (or countries/languages groups) analyzed in each round, please see: http://bit.ly/hH07b7

Differences between experiments may be observed: the media use experiment seems particularly problematic in both rounds, with no PS at all.

As seen in section 1, Saris et al. (2004) mention that in some cases the 2-group design may not be empirically identified, in particular when there is no correlation between the traits. This is what seems to happen in the media use experiment. The correlations between the reported time spent watching television, listening to the radio and reading newspapers are almost zero. This may explain the problems encountered. For the other topics however, the results are worse than expected from the reading of Saris et al. (2004) and there is no clear explanation. In addition, for the same experiment, sometimes within one country from one language to another, the SB-MTMM experiment may in one case provide directly a PS but in the other not.

The next section wants to investigate the consequential question: why? Under which conditions are the NC and HC occurring? Understanding when they are encountered may help finding how to solve them by preventing these conditions to happen. Based on the warnings made by Saris et al. (2004), four main explanations are considered:

- the role of the number of split-ballot groups: are there more problems in the 2-group SB design because of the incomplete design?
- the sample size: are the problems solved if the sample size is large enough?
- the closeness of the true values to boundaries: are the HC occurring because the true values are close to zero?
- the similarities between different true values: are there more non convergence problems because of these similarities?

3. Looking for explanations using Monte Carlo simulations

In order to shed some light on these questions, a Monte Carlo experiment is performed. In that way, different conditions can be tested and their (lack of) performance in terms of NC and HC is investigated.

3.1. The two cases of departure

In order to simulate data, two cases are selected: one "problematic case" (referred as case 1 from now on) is inspired by the satisfaction experiment studied in section 2.1[6], and one "working well case" (case 2) is inspired by values used in Saris et al. (2004) since the paper presents encouraging results even for the 2-group design. The values of the parameters in cases 1 and 2 are presented in Table 2. LISREL notation is used. The meaning of the symbols can be found in Appendix 1. The loadings associated with the method factors are not mentioned since they are all 1 according to our basic specification of the MTMM model. The same holds for the loadings between true scores and observed variables.

Table 2: List of values of the parameters

So far, the multiple-group (MG) approach of LISREL has been used. However, M*plus* offers a handier way of carrying out Monte Carlo simulations. Therefore, all the simulations are conducted using the Full Information Maximum Likelihood (FIML) of M*plus* 5.2 (Muthén and Muthén, 1998-2007). Both procedures are equivalent asymptotically, but with limited sample sizes differences may exist. An investigation of the consequences[7] of this change suggests that if the ESS data considered in section 2 would have been analysed with FIML instead of MG estimation, the frequency of problems encountered would have been even higher. Therefore, changing the estimation

[6] In order to have population values corresponding to a PS, instead of doing a SB-MTMM, we combined the matrices for groups 1 and 2, and filled in the part missing in both groups, trying to keep a similar pattern as for the rest of the matrix. Then, we estimated a 1-group MTMM. The values chosen for case 1 are based on these estimates.

[7] More details and the results can be found online: http://bit.ly/g3mFuA

procedure would not reduce the problems occurring in practice. On the other hand, using FILM estimation in the simulation studies may lead to more pessimistic results than the ones that would be encountered using the MG procedure.

3.2. The role of the number of SB groups

Saris et al. (2004) mention possible empirical non-identification for the 2-group but not for the 3-group design, which suggests that at least part of the problematic cases could be problematic only if a 2-group design is used, but could be solved if a 3-group design was implemented instead. So our first hypothesis is:

H1*: The frequency of NC and HC is higher in a 2- than in a 1- or 3-group design.*

In order to test this hypothesis, Monte Carlo studies are done using the two sets of values (cases 1 and 2). Based on the number of observations of the ESS, we decided to use 500 observations in each replication. Most of the ESS groups have more observations, but the idea was to take empirical conditions that are not favourable, since we want to study when problems are occurring. 500 datasets are generated according to the MTMM model in a 1-group (i.e. no split-ballot), 2-group and 3-group designs. The 1-group design is used as a reference point to judge the performance of other designs. Groups are approximately of similar sizes: for the 2-group design around 250 observations each, and for the 3-group around 170. An example of M*plus* code can be found online[8].

The number of convergent replications out of 500 is given in Table 3, together with the average estimates over all the convergent replications. Their standard deviations are mentioned in brackets. For the sake of clarity, we report only the results for six parameters: three loadings (γ_{11}, γ_{41}, γ_{71}) and three variances (φ_{44}, φ_{55} and φ_{66}). The

[8] Example of M*plus* code available here: http://bit.ly/gNMqEe

pattern for the other parameters is similar. Considering them would not change the trends observed. The table also provides indications about the "average bias" (sum of the absolute differences between the average estimates and the true values for the six parameters considered in the table divided by six) and the average mean squared errors (MSE)[9].

Table 3: Number of convergent replications and average estimates for different numbers of groups

It is clear from Table 3 that *H1* is confirmed: non convergence is only occurring in the 2-group design. It is much more present in case 1 than in case 2, as expected since case 1 was chosen to be problematic. But even in case 2, in the 2-group design, 21 out of the 500 replications (i.e. 4.2%) do not lead to convergence. Besides, when convergence is achieved, the estimates of the 2-group design are in average more biased and their standard deviations larger. In case 2, the 2-group design is performing relatively well, even though the 1- and 3-group designs perform better. But in case 1, the 2-group design is leading to an improper solution. On average, φ_{44} is negative, such that the average case can be seen as a Heywood case. Also, the standard errors are very large.

The 3-group design on the contrary performs well, not only in case 2, but also in case 1 (designed to be problematic). All replications converge and the average estimates reproduce accurately the true values. Even if the standard deviations are larger than for the 1-group design, they are all quite small. This suggests that the frequency of the problems encountered in practice when analysing real data may come from the use of the 2-group design. This is coherent with the findings of Saris et al. (2004). However, the authors presented the problem as if it was marginal ("under these rather elementary conditions", p. 333), giving the impression that the 2-group designs should perform well

[9] The number of Heywood cases is not considered because we had no handy way of counting them. We would have needed to consider each replication separately since in the Mplus output of the Monte Carlo simulations we do not get indications about how many replications have improper estimates.

most of the time, whereas Table 1 suggests that using real data, it performs often poorly. Table 3 shows that even with simulated data, the 2-group design leads to problems in quite some cases. Having only two SB groups seems to be a condition under which problems appear, whereas having a complete MTMM or a SB-MTMM with three SB groups seem to be enough for the estimation to work well. Therefore, the next section will focus on the 2-group design.

3.3. Under which conditions does the 2-group SB-MTMM design not work well?

3.3.1. When the sample size is small?

We have to notice that all the simulations were done for a sample size of 500. We anticipate this choice of a quite small sample size to have an impact on the results. We can expect that with larger sample sizes, fewer problems will arise and that with sample sizes large enough, both cases 1 and 2 will perform well. Case 2 already performs well with 500 observations. Case 1 does not. Table 4 shows what happens if we increase the sample size in case 1.

Table 4: Increasing the sample size for case 1

As expected, the number of convergent replications increases with the sample size. At the same time, the average bias and MSE decrease. But 10.000 to 15.000 observations are needed in order to get reasonable convergence and average bias (i.e. similar as the ones we get with 500 observations in case 2). If one can achieve such a sample size, the analyses will not be problematic anymore. But in practice it is generally not possible to collect so many observations.

Besides, even if the 2-group design always performs worse than the 1 or 3-group designs, there is a big difference between its performance for cases 1 and 2 that cannot

be explained by the sample size since it is kept fixed (Table 3). So focusing on the situation of 500 observations, we can wonder why problems do arise in case 1 much more than in case 2. What is differentiating both cases that could explain that the 2-group design works much better in case 2 than in case 1? One difference lies in the set of values chosen. Looking at Table 2, we can distinguish two main aspects that vary between cases 1 and 2 and could explain their difference in performance: the absolute and relative values of parameters.

3.3.2. **When the absolute values are close to a boundary?**

By absolute values, we mean the order of magnitude of the values: how close to the boundaries of possible values (e.g. 0 or 1) they are. Our hypothesis is:

H2*: The closer the true value of a parameter is to a boundary, the higher is the probability to get an improper estimate, just because of sampling fluctuations.*

This idea that because of sampling fluctuations Heywood cases may appear more easily for values close to the boundaries is logical and has already been studied a long time ago (e.g. Van Driel, 1978). For instance, if the true method variance is .05, then the estimated value will more probably be negative than if the true value is .40.

Another example is suggested by Saris et al. (2004): they mentioned that the 2-group design is not empirically identified if there is no correlation between the traits. This is again a question of absolute values: if the correlation between the traits is zero, the model is not identified, and if it is very close to zero, one may also expect problems for being at the border of a problematic point.

Comparing cases 1 and 2, we see that the method variances are lower in case 1, where the HC appear. However, many elements vary between the cases, so it is not possible to draw conclusions only from this comparison. Therefore, we use case 1

values, keeping as many as possible of the other parameters similar, but increasing the method variances to see if fewer problems are occurring. Alternatively, we reduce the method variances in case 2 to see if it leads to more problems. The values of other kinds of parameters could be chosen closer and closer to a boundary. But Table 3 shows that M_1 variance estimate is negative in average in case 1 and this is something often happening in analyses of real data, so we focus on the method variances.

We found no support for *H2*. On the one hand, increasing the true values of the method variances in case 1 does not solve the non convergence problems. It also does not substantially reduce the average bias. Besides, on average, the estimated M_1 variance (φ_{44}) is still negative in case 1 when its true value is increased to .20. On the other hand, reducing the true values of the method variances in case 2 does not lead to more problems. The number of convergent replications is similar in all three situations and the estimates are accurate for all three too. Even when the true method variances are .05, the average bias is only .0041[10]. We have to be careful about generalising the results since we only varied the values of the method variances in this analysis, but it seems that the mechanisms in hypothesis 2 are not explaining differences between case 1 and case 2.

3.3.3. **When some relative values are too similar?**

By relative values, we mean the values of one parameter compared to the values of other parameters in the model. Saris et al. (2004) warn that if the correlations between the traits are all equal then the model is not empirically identified. Besides, we can see that in case 1 ("problematic case"), the correlations between the traits are more similar than in case 2 ("working well case"). All this suggests framing the null hypothesis in

[10] The table with the results is available online: http://bit.ly/eGRNiC

that way: the more similar the true values of the parameters, the higher the probability of facing problems.

However, the similarity of other parameters could as well play a role. Indeed, it is important to remark that case 1 has more variation in relative values of the parameters than case 2 (except for the correlation between traits): in case 2, the loadings are the same for the nine true scores, the error terms and method variances too (cf. Table 2), whereas in case 1 most parameters have different values. This goes against the logic just proposed. It may be necessary to be more precise. Similarity between some kinds of parameters may help the estimation, whereas similarity of others may on the contrary provoke problems.

How the similarity of values of parameters affects the performance of the 2-group design is therefore difficult to predict. A huge number of different patterns of variations (some parameters have the same true values, others not) can be considered. To determine if similarity of the true values of some parameters impacts the performance of the 2-group design, we focus on eight main conditions summarised in Table 5.

Table 5: The 8 conditions considered

Much more conditions could be studied, for instance if $\gamma_{11}=\gamma_{22}$, but both are different from γ_{33}. We did not formulate hypotheses for the eight conditions, because we do not have enough information to predict what could happen. But following Saris et al. (2004), we propose one hypothesis for the correlations between the traits:

H3*: The more similar the true correlations between the traits, the higher the probability of facing problems of NC or improper solutions.*

In order to see if the different conditions have an impact on the performance of the 2-group design, we use the values of case 2 as starting point to complete the other conditions' values. It corresponds to condition 5 in the table. The same values are used in the different conditions when the same equalities hold. For instance, conditions 5 to 8

share the fact that the true correlations values between the traits are different: in these four conditions, the value of the correlation between traits 1 and 2 is therefore the same (.60). All values for the eight conditions are provided in Table 6 ("first set of values")[11].

Table 6: First set of values for the 8 conditions

For each condition, datasets of 500 observations are generated and analysed. The number of replications converging out of the 500 and the average estimates for the convergent replications are given in Table 7. The true values can be found in Table .

Table 7: Results for the 8 conditions

What we see in Table 7 is the contrast between conditions where the values of the correlations between different pairs of traits are similar and conditions where they are not. Besides this expected difference (cf. Saris et al., 2004), there is no much variation between the conditions. We can notice also that for example in conditions 3 and 4, in average, proper solutions are obtained for the convergent replications. Nevertheless, the bias of the estimates is quite large. The results of real data analyses may be even more problematic than suggested by Table 1: even if a PS is achieved, this does not guarantee that the estimates are accurate. Table 7 shows they can indeed be quite biased. So we should be cautious also when a PS is directly obtained, at least with the 2-group design.

Table 7 suggests the logic of similarity between the parameters leading to different performance levels only holds for the correlation between traits. However, so far we only considered the cases where the correlations are equal or different. In order to test *H3*, we define sets of values for the correlations between traits, departing from the previous example, and varying the level of similarities. Remembering *H2* and the warning in Saris et al. (2004) about the non identification of the 2-group design if the correlations between the traits are zero, we also vary the absolute values. The 14 sets of values are presented in the left part of Table 8. The column "sum diff" gives the sum of

[11] As always: $\gamma_{14}= \gamma_{24}= \gamma_{34}= \gamma_{45}= \gamma_{55}= \gamma_{65}= \gamma_{76}= \gamma_{86}= \gamma_{96}=1$ and $\lambda_{11}= \lambda_{22}=...=\lambda_{99}=1$ and $\varphi_{11}= \varphi_{22}= \varphi_{33}=1$

the absolute value of the differences in true absolute values between the correlations: *sumdiff* = | $|\varphi_{21}|$- $|\varphi_{31}|$ | + | $|\varphi_{21}|$-| φ_{32}| | + | $|\varphi_{31}|$- $|\varphi_{32}|$ |. According to *H3*, the smaller *sumdiff,* the higher the risk of getting problems.

Each set of values is used to do a Monte Carlo simulation for the four different conditions with unequal true values of the correlations between traits (conditions 5 to 8), leading to 14*4 = 56 simulations and a huge amount of results. The number of convergent replications out of 500 is indicated in Table 8, together with the average bias for the six parameters considered so far and the average MSE. However, for the sake of clarity, the estimates are only available online[12]. Also, a simplistic judgement on the performance for each set of values is proposed, based on the number of convergent replications, the bias of the estimates and their standard deviations: "++"means the performance is good, "+" that it is quite good, "-" that it is quite poor and "- -" that it is poor.

Table 8: results for the different sets of values and conditions

One can see in Table 8 that the number of convergent replications varies across sets of values: for example in condition 7, it goes from 272 (14th set) to 500 (9th set). Across conditions also some variations are observed, even if a bit more limited. In the 13th set of values, 290 replications converge in condition 6 against 416 in condition 8. Said differently, there are 126/500 = 25.2% more convergent replications in condition 8 than in condition 6.

For the convergent replications, the bias can be considered. Again, variability appears both across sets of values and conditions. Condition 8, where none of the true values are equal, performs a bit better in general. On the contrary, condition 5, where all gammas have the same true value, leads to more NC and less accurate estimates.

[12] Please, see: http://bit.ly/kWK1M2

But is *H3* verified? There is some evidence going in this direction: the sets of values with the lowest "sum diff" are all performing quite badly. However, other elements seem to play a role. Indeed, the 8th set of values performs poorly even if its "sum diff" is large. On the contrary, the 4th and 5th perform quite well even with small "sum diff".

So there are probably more than one element explaining that one case works well and another not. A mixture of both the absolute and relative values would be coherent with results in Table 8. We can think also that even if the "sum diff" is high, if at least two traits have a low correlation, then the probability of getting undesirable results is higher.

In order to investigate these points further, first the number of convergent replications and then the average bias presented in Table 8 are used as dependent variables in regression analyses. Several independent variables are included, according to the different hypotheses mentioned: dummies for the different conditions (condition 5 is taken as reference), sum of the differences between the trait correlations, minimum absolute correlation between two of the traits (*mincorrphi*= $min[|\varphi_{21}|,|\varphi_{31}|,|\varphi_{32}|]$), smallest difference between two of the correlations between the traits (*smalldiff* = $Min(|\ |\varphi_{21}|-|\varphi_{31}|\ |,\ |\ |\varphi_{21}|-|\varphi_{32}|\ |,\ |\ |\varphi_{31}|-|\varphi_{32}|\ |)$), and interactions between the minimum correlation and the sum or the smallest difference. Table 9 presents the coefficients of these two regressions.

Table 9: Coefficients of the regressions of number of convergent replications and average bias

Table 9 shows that the smallest difference and the interaction between the sum of the differences and the minimum correlation have a significant impact on the number of convergent replications: the higher they are, the higher the convergence. Looking at the average bias, the smallest difference does not reach significance, but the interaction between sum difference and minimum correlation does: the higher this interaction, the

lower the average bias. Condition 8 also reaches significance: the average bias is lower in this condition where all gammas have different true value.

We can notice that some of the effects estimated are very large: for instance, even if the maximum number of convergent replications in our replications is 500, the impact of *smalldiff* on the number of convergent replications is almost 622 with a constant of 302. But the variable *smalldiff* takes values between 0 and .5 so the increase in the number of convergent replications at the end is not so high. For the interaction terms, the impact is even smaller since we have to multiply also by the variable *mincorrphi* that is in the [0,1] interval. Besides, if we consider the regression equation for the number of convergent replications, we can see that we have not only large positive effects but also one large negative effect (*smalldiff*mincorrphi*), even if the latter is not significant.

As a last check, we try to see if the results we got using simulations seem to apply to our real data. Table 1 contains information about the number of non convergence in different SB-MTMM experiments collected as part of the ESS. We already mentioned that the media experiment is probably problematic because of the lack of correlation between the traits. The regression suggests that there are also more non-convergence problems when the interaction between *sumdiff* and *mincorrphi* is low. Of course, for real data we do not know the true correlations between the traits, but we can still try to estimate *sumdiff* and *mincorrphi* based on values obtained in some analyses[13]. Doing so, we found that indeed, the real data experiments with more non convergence problems were the same as the ones with lower *sumdiff*mincorrphi* (Appendix 2). We also looked at the relation with *smalldiff*, but in that case the link is less clear, which may be relate to the fact that the smallest difference is always very low: we do not have

[13] We used estimates for ph21, ph31 and ph32 obtained by applying the SB-MTMM true score model in a multiple-group analysis with all the ESS countries for a given round included. These estimates can still be biased but we used them as an approximation we had of what the true correlations could be.

therefore many variations in this variable, which might be why it is not possible to see any clear effect.

4. Conclusions

The SB-MTMM approach proposed by Saris et al. (2004) is attractive because it combines the advantages of two well-known approaches. It reduces the cognitive burden of the respondents and allows using MTMM experiments in shorter surveys. However, the implementation of the SB-MTMM approach has led to frequent problems of non convergence and Heywood case. Quite some difficulties have been encountered trying to analyse the existing data, in particular the 2-group SB-MTMM experiments of the ESS. Therefore, more attention needed to be paid to this interesting but still problematic design. This was the motivation for this study. One main question guided our analyses: under which conditions is the SB-MTMM approach performing well and under which conditions is it not?

The paper has obvious limits. A lot of choices had to be made: many more conditions, sets of values, etc, could be studied. Besides, all simulations are done using the correct model to generate and analyse the data. More work would therefore be needed to test the robustness of our results and to see what is happening when using a slightly misspecified model to analyse our data as it is most of the time happening when analysing real data.

However, overall, the Monte Carlo simulations conducted suggest that problems occur with the 2-group design but not with the 3-group design. The number of SB groups used is the first main condition determining if the SB-MTMM approach is or is not performing well. This result is in line with Saris et al. (2004), who mention that the

2-group design may not be empirically identified in some cases, whereas the 3-group design normally is. However, they present the conditions of empirical non identification as quite unlikely, whereas the real data analyses show frequent problems.

Nevertheless, the 2-group design does not perform systematically poorly: there are cases where it performs well and others where it does not. If the sample size is large enough, even problematic cases appear to perform quite well in the 2-group design. But "large enough" sometimes means between 10.000 and 15.000 observations. This is in practice much too high for many researchers. Besides, for 500 observations, big differences in the performance of the 2-group design can be observed. So focusing on the 2-group design with 500 observations, two hypotheses trying to explain these differences were tested: the closer the absolute true values to boundaries, the higher the probability of getting problems; and the more similar the true correlations between the traits, the higher the probability of getting problems. The second received more support, but regression analyses suggested that the interaction between the absolute true values of the correlation between the traits and differences in correlations between the traits has a significant effect on the convergence and on the bias. So complex mechanisms are at work to determine when the 2-group design performs properly.

In order to avoid problems in practice, we therefore recommend using a 3-group design as often as possible when designing a SB-MTMM experiment. Moreover, the traits should be chosen is such a way that they are correlated with each other, but not to the same extent. It may be however quite difficult in practice to know in advance which traits are satisfying these conditions. Pre-tests on small samples might help to design SB-MTMM experiments that do not convey too many non convergence and Heywood cases problems.

References

Alwin, D.F. (1997). "Feeling Thermometers Versus 7-Point Scales: Which are Better?" *Sociological Methods and Research* 25 (3):318.

Alwin, D.F. (2007). *Margins of errors: a study of reliability in survey measurement*. Wiley-Interscience

Andrews, F. (1984). "Construct validity and error components of survey measures: A structural modeling approach." *Public Opinion Quarterly*, 46, 409-42. Reprinted in W.E. Saris & A. van Meurs. (1990). *Evaluation of measurement instruments by metaanalysis of multitrait multimethod studies*. Amsterdam: North-Holland

Browne, M.W. (1984). "The decomposition of multitraitmultimethod matrices". *British Journal of Mathematical and Statistical Psychology*, 37, 1-21.

Campbell, D.T. and Fiske, D.W. (1959). "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 6:81-105

Corten, I.W., Saris, W.E., Coenders, G.M., van der Veld, W., Aalberts, C.E., and Kornelis, C. (2002). "Fit of different models for multitrait-multimethod experiments". *Structural Equation Modeling*, 9(2), 213-232.

Cudeck, R. (1988). "Multiplicative Models and MTMM Matrices." *Journal of Educational Statistics* 13 (2):131-147.

Eid, M. (2000). "A Multitrait-Multimethod Model with Minimal Assumptions." *Psychometrika* 65 (2):241-261.

Jöreskog, K.G. (1970). "A general method for the analysis of covariance structures". *Biometrika*, 57:239-51

Jöreskog, K.G. (1971). "Statistical analysis of sets of congeneric tests". *Psychometrika*, 36, 109-133.

Jöreskog, K.G. and Sörbom, D. (1991). *LISREL VII: A guide to the program and applications*. Chicago, IL: SPSS.

Kolenikov, S., and K.A. Bollen. (2008). "Testing Negative Error Variances: Is a Heywood Case a Symptom of Misspecification?" *University of North Carolina*

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

Muthén, L.K. and Muthén, B.O. (1998-2007). Mplus User's Guide. Fifth Edition. Los Angeles, CA: Muthén & Muthén

Revilla, M., Saris, W.E., and J.A. Krosnick (2009). "Choosing the number of categories in agree-disagree scales" RECSM Working Paper 5.

Rindskopf, D. (1984). "Structural Equation Models." *Sociological Methods & Research* 13 (1):109.

Saris, W.E., and C. Aalberts (2003). "Different Explanations for Correlated Disturbance Terms in MTMM Studies." *Structural Equation Modeling: A Multidisciplinary Journal* 10 (2):193-213.

Saris, W.E. and Andrews, F.M. (1991). "Evaluation of measurement instruments using a structural modeling approach". In P.P. Biemer, R.M. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (Eds.), *Measurement errors in surveys* (pp. 575-597). New York: Wiley.

Saris, W.E. and Gallhofer, I. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. New York:Wiley

Saris, W.E., Satorra, A. and Coenders, G. (2004). "A new approach to evaluating the quality of measurement instruments: The split-ballot MTMM design." *Sociological Methodology 2004*

Saris, W.E, Satorra, A., Van der Veld, W.M. (2009). "Testing Structural Equation Models or Detection of Misspecifications?" *Structural equation modeling: A multidisciplinary Journal*, 16 (4): 561-582

Scherpenzeel, A.C. (1995). "A question of quality: Evaluating survey questions by multitrait-multimethod studies". Amsterdam, Nimmo.

Van der Veld, W.M., Saris, W.E., Satorra, A. (2008) Judgment Aid Rule Software

Van Driel, O. P. (1978). "On various Causes of Improper Solutions in Maximum Likelihood Factor Analysis." *Psychometrika* 43 (2):225-243.

Van Meurs, L. and Saris, W.E. (1990). Memory effects in MTMM studies. In W.E. Saris and L. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies*. Amsterdam: North Holland.

Werts, C.E., and Linn, R.L. (1970). Path analysis: Psychological examples. *Psychological Bulletin*, 74, 194-212.

Wothke, W.. (1996). "Models for Multitrait-Multimethod Matrix Analysis." *Advanced Structural Equation Modeling: Issues and Techniques* 7-56.

Appendix 1: Notations in Lisrel

γ = gamma = ga λ = lambda y = ly θ = teta epsilon = te φ = phi = ph	ga11 = loading between T_{11} and F_1 ga22 = loading between T_{21} and F_2 ga33 = loading between T_{31} and F_3 ga41 = loading between T_{12} and F_1 ga52 = loading between T_{22} and F_2 ga63 = loading between T_{32} and F_3 ga71 = loading between T_{13} and F_1 ga82 = loading between T_{23} and F_2 ga93 = loading between T_{33} and F_3
te11 = error term associated to Y_{11} te22 = error term associated to Y_{22} te33 = error term associated to Y_{33} te44 = error term associated to Y_{44} te55 = error term associated to Y_{55} te66 = error term associated to Y_{66} te77 = error term associated to Y_{77} te88 = error term associated to Y_{88} te99 = error term associated to Y_{99}	ga14 = loading between T_{11} and M_1 ga24 = loading between T_{21} and M_1 ga34 = loading between T_{31} and M_1 ga45 = loading between T_{12} and M_2 ga55 = loading between T_{22} and M_2 ga65 = loading between T_{32} and M_2 ga76 = loading between T_{13} and M_3 ga86 = loading between T_{23} and M_3 ga96 = loading between T_{33} and M_3
ph11 = trait 1 variance ph22 = trait 2 variance ph33 = trait 3 variance ph44 = method 1 variance ph55 = method 2 variance ph66 = method 3 variance ph21 = correlation between traits 1 and 2 ph31 = correlation between traits 1 and 3 ph32 = correlation between traits 2 and 3	ly11 = loading between T_{11} and Y_{11} ly22 = loading between T_{21} and Y_{21} ly33 = loading between T_{31} and Y_{31} ly44 = loading between T_{12} and Y_{12} ly55 = loading between T_{22} and Y_{22} ly66 = loading between T_{32} and Y_{32} ly77 = loading between T_{13} and Y_{13} ly88 = loading between T_{23} and Y_{23} ly99 = loading between T_{33} and Y_{33}

Appendix 2:

Round	Experiment	NC	*sumdiff*mincorrphi*	*mincorrphi*	*smalldiff*	|Ph21|	|Ph31|	|Ph32|
4	Political trust	1	.2184	.42	.03	.65	.42	.68
1	Political efficacy	1	.2142	.51	.04	.55	.72	.51
1	Political trust	2	.2052	.38	.06	.59	.38	.65
1	Satisfaction	3	.1920	.60	.06	.76	.60	.66
1	Social trust	3	.1584	.66	.02	.76	.66	.78
1	Political orientation	4	.0420	.07	.01	.08	.07	.37
4	Satisfaction	9	.1980	.66	.01	.81	.80	.66
1	Media	15	.0028	.01	.07	.08	.15	.01
4	Media	16	.0060	.05	.02	.05	.11	.07

Note: NC = number of non convergent cases

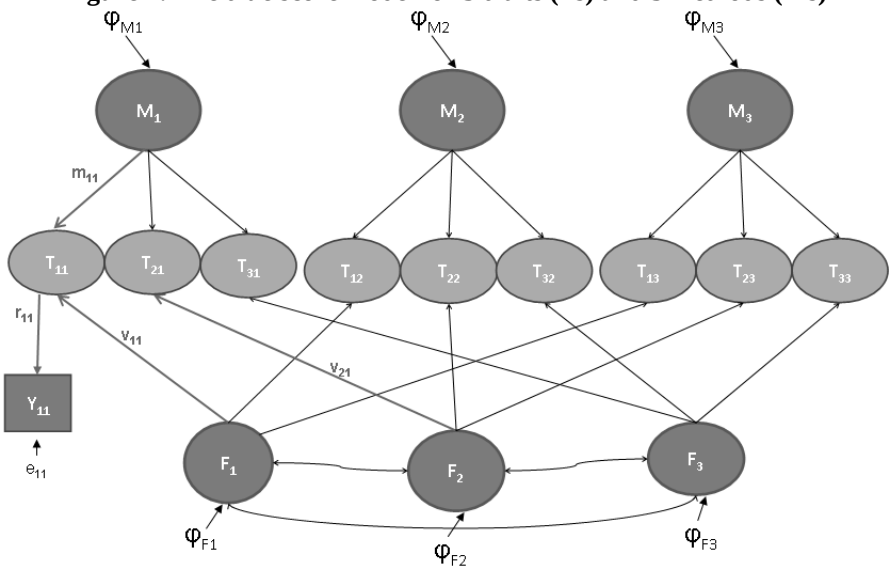
sumdiff=||*ph21*|-1*ph31*|| + ||*ph21*|-|*ph32*|| + ||*ph31*|-|*ph32*||

mincorrphi= Min(|*ph21*|,|*ph31*|,|*ph32*|*);*

smalldiff= Min(||*ph21*|-|*ph31*||, ||*ph21*|-|*ph32*||, ||*ph31*|-|*ph32*||*)*

Figures

Figure 1: The true score model for 3 traits (Fs) and 3 methods (Ms)



Tables

Table 1: Results obtained when running 180 SB-MTMM models for ESS rounds 1 and 4

Experiments		NC	HC	PS	Total cases
Round 1	Media use	15	4	0	19
	Pol. efficacy	1	11	7	19
	Pol. orientation	4	8	7	19
	Satisfaction	3	9	7	19
	Social trust	3	13	3	19
	Political trust	2	10	7	19
Round 4	Media use	16	6	0	22
	Satisfaction	9	10	3	22
	Political trust	1	13	8	22
Total across experiments (Total in %)		54 (30.0%)	84 (46.7%)	42 (23.3%)	180 (100%)

Note: NC = not convergent, HC = Heywood case, PS = proper solution

Table 2: List of values of the parameters

	Case1	Case2		Case1	Case2		Case1	Case2
Ga11	.74	.735	Te11	.35	.30	Ph21	.46	.60
Ga22	.82	.735	Te22	.23	.30	Ph31	.50	.10
Ga33	.74	.735	Te33	.35	.30	Ph32	.43	.30
Ga41	.70	.735	Te44	.45	.30	Ph11	1	1
Ga52	.74	.735	Te55	.39	.30	Ph22	1	1
Ga63	.74	.735	Te66	.39	.30	Ph33	1	1
Ga71	.86	.735	Te77	.17	.30	Ph44	.10	.16
Ga82	.86	.735	Te88	.17	.30	Ph55	.06	.16
Ga93	.83	.735	Te99	.22	.30	Ph66	.09	.16

Table 3: Number of convergent replications and average estimates for different numbers of groups

	Case 1				Case 2			
500 obs	1-group	2-group	3-group	true	1-group	2-group	3-group	true
Ga11	.7386	.9838	.7384	.74	.7341	.7423	.7335	.735
(SD)	(.0429)	(.4230)	(.0592)		(.0432)	(.0663)	(.0605)	
Ga41	.6998	.5840	.6994	.70	.7336	.7277	.7318	.735
(SD)	(.0425)	(.1680)	(.0600)		(.0418)	(.0753)	(.0607)	
Ga71	.8571	.7190	.8590	.86	.7315	.7277	.7342	.735
(SD)	(.0393)	(.1987)	(.0583)		(.0421)	(.0760)	(.0599)	
Ph44	.1001	**-.2007**	.1004	.10	.1607	.1562	.1614	.16
(SD)	(.0186)	(.6787)	(.0279)		(.0238)	(.0377)	(.0336)	
Ph55	.0599	.1263	.0591	.06	.1594	.1605	.1581	.16
(SD)	(.0174)	(.0875)	(.0282)		(.0228)	(.0441)	(.0351)	
Ph66	.0888	.1771	.0865	.09	.1585	.1579	.1554	.16
(SD)	(.0176)	(.1153)	(.0329)		(.0238)	(.0453)	(.0357)	
Number conv	500	**266**	500	500	500	479	500	500
Average bias	.0010	**.1592**	.0013		.0014	.0047	.0022	
Average MSE	.0315	**.1771**	.0327		.0428	.0452	.0439	

Table 4: Increasing the sample size for case 1

2-group	Number of observations										
Case 1	500	800	1000	1500	2000	5000	7500	10000	15000	20000	true
Ga11	.9838	.9619	.9494	.8998	.8880	.8080	.7672	.7546	.7480	.7448	.74
(SD)	(.4230)	(.4331)	(.4032)	(.3734)	(.2996)	(.2171)	(.1151)	(.0792)	(.0690)	(.0584)	
Ga41	.5840	.5998	.6029	.6255	.6215	.6653	.6868	.6935	.6982	.7001	.70
(SD)	(.1680)	(.1638)	(.1601)	(.1487)	(.1354)	(.1098)	(.0852)	(.0701)	(.0611)	(.0548)	
Ga71	.7190	.7424	.7466	.7767	.7688	.8179	.8449	.8534	.8584	.8606	.86
(SD)	(.1987)	(.2005)	(.1955)	(.1815)	(.1667)	(.1327)	(.1024)	(.0845)	(.0731)	(.0647)	
Ph44	-.2007	-.1813	-.1581	-.0973	-.0664	.0224	.0726	.0859	.0919	.0951	.10
(SD)	(.6787)	(.7417)	(.6435)	(.6485)	(.4189)	(.3372)	(.1131)	(.0647)	(.0557)	(.0458)	
Ph55	.1263	.1150	.1116	.0983	.1026	.0773	.0654	.0620	.0595	.0586	.06
(SD)	(.0875)	(.0895)	(.0876)	(.0832)	(.0770)	(.0653)	(.0559)	(.0489)	(.0430)	(.0382)	
Ph66	.1771	.1617	.1595	.1411	.1466	.1130	.0967	.0919	.0887	.0875	.09
(SD)	(.1153)	(.1234)	(.1205)	(.1136)	(.1049)	(.0910)	(.0765)	(.0668)	(.0583)	(.0527)	
Number conv	266	281	302	324	340	410	450	462	492	496	500
Average bias	.1592	.1293	.1216	.0922	.0878	.0400	.0147	.0073	.0033	.0020	
Average MSE	.1771	.1872	.1573	.1442	.0937	.0833	.0387	.0347	.0334	.0325	

Table 5: The 8 conditions considered

Conditions	1	2	3	4	5	6	7	8
Ph21=ph31=ph32	Yes	Yes	Yes	Yes	No	No	No	No
Ga11=ga22=ga33 Ga41=ga52=ga63 Ga71=ga82=ga93 Te11=te22=te33 Te44=te55=te66 Te77=te88=te99	Yes	Yes	No	No	Yes	Yes	No	No
Ga11=ga41=ga71 Ga22=ga52=ga82 Ga33=ga63=ga93 Te11=te44=te77 Te22=te55=te88 Te33=te66=te99	Yes	No	No	Yes	Yes	No	Yes	No
Ph44=ph55=ph66								

Note: "yes" = the equalities of the true values of the parameters indicated in the first column are specified in one condition; "no" = none of the parameters in a given equality raw are equal. For instance, in condition 3, γ_{11} is different from γ_{22} and from γ_{33}, but also γ_{22} is different from γ_{33}. In that condition also γ_{11} is different from γ_{41} and for γ_{71}, and γ_{41} is different from γ_{71}. In fact, all gammas are different in condition 3.

Table 6: First set of values for the 8 conditions

	Conditions									Conditions							
	1	2	3	4	5 =case2	6	7	8		1	2	3	4	5	6	7	8
Ph21	.30	.30	.30	.30	.60	.60	.60	.60	Ph44	.16	.20	.20	.16	.16	.20	.16	.20
Ph31	.30	.30	.30	.30	.10	.10	.10	.10	Ph55	.16	.10	.10	.16	.16	.10	.16	.10
Ph32	.30	.30	.30	.30	.30	.30	.30	.30	Ph66	.16	.16	.16	.16	.16	.16	.16	.16
Ga11	.735	.70	.70	.70	.735	.70	.70	.70	Te11	.30	.31	.31	.35	.30	.31	.35	.31
Ga22	.735	.70	.75	.75	.735	.70	.75	.75	Te22	.30	.31	.24	.28	.30	.31	.28	.24
Ga33	.735	.70	.85	.85	.735	.70	.85	.85	Te33	.30	.31	.08	.12	.30	.31	.12	.08
Ga41	.735	.80	.80	.70	.735	.80	.70	.80	Te44	.30	.26	.26	.35	.30	.26	.35	.26
Ga52	.735	.80	.70	.75	.735	.80	.75	.70	Te55	.30	.26	.41	.28	.30	.26	.28	.41
Ga63	.735	.80	.75	.85	.735	.80	.85	.75	Te66	.30	.26	.34	.12	.30	.26	.12	.34
Ga71	.735	.65	.65	.70	.735	.65	.7	.65	Te77	.30	.42	.42	.35	.30	.42	.35	.42
Ga82	.735	.65	.80	.75	.735	.65	.75	.80	Te88	.30	.42	.20	.28	.30	.42	.28	.20
Ga93	.735	.65	.70	.85	.735	.65	.85	.70	Te99	.30	.42	.35	.12	.30	.42	.12	.35

Table 7: results for 8 conditions

Conditions	1	2	3	4	5	6	7	8
Ga11	1.0265	.9998	.8607	.8876	.7423	.7072	.7063	.7049
(SD)	(5200)	(.5251)	(.3651)	(.2652)	(.0663)	(.0672)	(.0639)	(.0626)
Ga41	.6055	.6520	.7210	.5854	.7277	.7923	.6921	.7954
(SD)	(.1925)	(.2174)	(.2028)	(.1584)	(.0753)	(.0802)	(.0748)	(.0806)
Ga71	.6031	.5302	.5797	.5867	.7277	.6433	.6920	.6438
(SD)	(.1924)	(.1880)	(.1702)	(.1602)	(.0760)	(.0818)	(.0759)	(.0764)
Ph44	**-.0629**	**-.0244**	.0765	.0418	.1562	.1963	.1570	.1967
(SD)	(.5546)	(.5185)	(.3244)	(.1847)	(.0377)	(.0385)	(.0328)	(.0400)
Ph55	.2017	.1520	.1241	.2041	.1605	.1005	.1608	.1005
(SD)	(.0722)	(.0846)	(.0722)	(.0651)	(.0441)	(.0472)	(.0388)	(.0431)
Ph66	.1975	.1892	.1780	.2000	.1579	.1574	.1587	.1580
(SD)	(.0732)	(.0666)	(.0678)	(.0660)	(.0453)	(.0456)	(.0399)	(.0441)
Number conv	**226**	**237**	**284**	**228**	479	481	466	492
Average bias	**.1425**	**.1455**	**.0793**	**.1030**	.0047	.0047	.0046	.0036
Average MSE	**.1719**	**.1689**	**.1023**	**.0941**	.0452	.0525	.0552	.0522

Table 8: results for the different sets of values and conditions

Characteristics sets of values					Number conv				Average bias				Average MSE				CCL
Set	Values for			Sum	conditions				conditions				conditions				
number	Ph21	Ph31	Ph32	diff	5	6	7	8	5	6	7	8	5	6	7	8	All
1	.60	.30	.10	1.00	479	481	466	492	.00	.00	.00	.00	.0452	.0525	.0552	.0522	+ +
2	.10	.20	.30	**.40**	348	352	367	383	.08	.08	.05	.04	.1168	.1120	.0813	.0748	- -
3	.40	.60	.20	.80	486	483	471	485	.00	.00	.00	.00	.0450	.0523	.0547	.0523	+ +
4	.40	.50	.20	.60	467	467	434	476	.01	.00	.01	.01	.0466	.0538	.0564	.0543	+ +
5	.30	.40	.10	.60	451	452	425	461	.01	.01	.01	.01	.0468	.0539	.0564	.0539	+ +
6	.20	.30	.40	**.40**	390	391	407	422	.05	.05	.03	.02	.0947	.0770	.0647	.0619	- -
7	.80	.90	.70	**.40**	444	500	429	468	.02	.03	.03	.01	.0542	.1256	.0642	.0585	-
8	.10	.20	.80	**1.40**	431	423	456	446	.03	.03	.01	.02	.0805	.0926	.0623	.0753	-
9	.85	.90	.50	.80	500	500	**500**	500	.00	.00	.00	.00	.0441	.0517	.0539	.0515	+ +
10	.10	.25	.40	**.60**	426	427	444	448	.03	.03	.02	.01	.0588	.0681	.0627	.0663	-
11	.20	.30	.90	1.4	469	474	487	476	.01	.01	.00	.00	.0496	.0575	.0569	.0545	+ +
12	.70	.20	.10	1.2	429	431	360	453	.01	.01	.01	.01	.0450	.0520	.0555	.0522	+
13	.40	.45	.50	**.20**	294	**290**	396	**416**	.12	.10	.04	.03	.2361	.1805	.0869	.0918	- -
14	.30	.20	.10	**.40**	363	362	**272**	366	.04	.04	.04	.04	.0650	.0736	.0636	.0685	- -

Note: "++"= performance is good, "+" = quite good, "-" = quite poor, "- -" = poor.
Sum diff in **bold** = conclusion about the performance is quite poor or poor

Table 9: Coefficients of the regressions of number of convergent replications and average bias on a set of explanatory variables

	Number conv	Average bias
Sumdiff	9.86	-.0015
Condition 6	4.00	-.0014
Condition 7	-4.50	-.0114
Condition 8	22.50	-.0150*
Mincorrphi	-22.19	.0926
Smalldiff	621.90*	-.1819
Sumdiff*mincorrphi (interaction)	418.83*	-.2090*
Smalldiff*mincorrphi (interaction)	-223.18	-.3132
Constant	301.95	.0660
Adjusted R^2	.6014	.5387
Number of observations = 56	P<.05 indicated by a star	

Note: sumdiff=||ph21|-1ph31|| + ||ph21|-|ph32|| + ||ph31|-|ph32||
mincorrphi= Min(|ph21|,|ph31|,|ph32|);
smalldiff= Min(||ph21|-|ph31||, ||ph21|-|ph32||, ||ph31|-|ph32||)
**: significant at 5%*