DE GRUYTER OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 117–135, http://dx.doi.org/10.2478/JOS-2019-0006

Cross-National Comparison of Equivalence and Measurement Quality of Response Scales in Denmark and Taiwan

Pei-shan Liao[1], *Willem E. Saris*[2], *and Diana Zavala-Rojas*[3]

The split-ballot multitrait-multimethod (SB-MTMM) approach has been used to evaluate the measurement quality of questions in survey research. It aims to reduce the response burden of the classic MTMM design, which requires repeating alternative formulations of a survey measure to the same respondent at least three times, by using combinations of two methods in multiple groups. The SB-MTMM approach has been applied to the European Social Survey (ESS) to examine the quality of questions across countries, including the differences in response design and measurement errors. Despite wide application of the SB-MTMM design in Europe, it is yet unknown whether the same quality of survey instruments can be achieved in both a different cultural context and in a logographic writing system, like the one in Taiwan.

This study tests for measurement invariance and compares measurement quality in Taiwan and Denmark, by estimating the reliability and validity of different response scales using the SB-MTMM approach. By using the same questions as in the ESS, a cross-cultural comparison is made, in order to understand whether the studied response scales perform equally well in Taiwan, compared to a European country. Results show that quality estimates are comparable across countries.

Key words: Split-ballot MTMM; reliability; validity; question quality.

1. Introduction

Survey measures take various forms, and studying their quality is an important issue, as they result in measurement error biases. For example, questions about subjective concepts can be measured by Likert-type response scales with different numbers and labels of response categories, with a feeling thermometer or using rating scores, among others (Alwin 1997; Schaeffer and Presser 2003). In comparative survey research, different measurement designs influence the response distributions and may lead to comparability problems across countries (see Bjørnskov 2010). Among the tools used to evaluate measurement quality, two approaches have been rather popular: the split-ballot experiment

[1] Center for Survey Research, RCHSS, Academia Sinica, 128 Academia Road, Sec. 2, Nangang Dist., Taipei 11529, Taiwan. Email: psliao@gate.sinica.edu.tw
[2] Sociometric Research Foundation, Carer Josep Pla 27 9-4, 08019 Barcelona, Spain. Email: w.saris@telefonica.net
[3] RECSM, Universitat Pompeu Fabra, Ramon Trias Fargas, 25-27, Edifici Mercè Rodoreda 24, 08005 Barcelona, Spain
Acknowledgments: This study was based on the Taiwan Social Change Survey, supported by the Ministry of Science and Technology, Taiwan (MOST 104-2420-H-001-005-SS3).

and the multitrait-multimethod (MTMM) approach (Saris et al. 2010; Saris et al. 2004). The basic principle of the split-ballot experiment approach is to randomly divide the respondents into two or more equal-sized subsamples with equal representativeness of the total sample (Schuman and Presser 1981; Petersen 2008). The respondents of each subsample answer survey questions simultaneously and under the same conditions. Variations in the questionnaire for each of the subsamples are treated as an experimental stimulus to examine questionnaire effects.

Alternatively, Campbell and Fiske (1959) suggested the MTMM design to evaluate the validity of social science concepts based on the correlations among measures of variables (Alwin 1974). The classic MTMM approach requires a respondent to answer questions about a minimum of three *traits,* that is, concepts or constructs measured using three different methods, for example, response scales, leading to nine different observed variables (Saris and Gallhofer 2014). Given the matrix, criteria for convergent and discriminant validity of these variables are advanced in Campbell and Fiske (1959) to assess validity. Structural equation modeling (SEM) can be applied to estimate the reliability and validity of each method. A comparison of fit statistics indicates which model best fits the data. Since the respondents need to repeatedly answer similar questions, it becomes a burden for them and may cause memory bias or order effect of the questions.

Saris et al. (2004) developed an approach to reduce the response burden by means of using different combinations of two methods in multiple groups. They combine the use of multiple groups in a split-ballot design, while the MTMM approach allows estimating the reliability and validity of the different questions. Such a split-ballot MTMM (SB-MTMM) approach has been applied to the European Social Survey (ESS) to examine the measurement quality of questions across countries, including the differences in response design and measurement errors (Oberski et al. 2007, 2010; Saris and Gallhofer 2014; Saris et al. 2008; Saris et al. 2010). Information about the quality of more than 2,700 questions from different European countries and the United States are stored in an online database in the Survey Quality Predictor (SQP) 2.1, which is an online system for survey quality prediction (Saris and Gallhofer 2014; Saris et al. 2011). On the basis of the data collected in all these countries using mainly English and European languages, a meta-analysis has been performed to develop a procedure to predict the quality of survey questions. This prediction tool is available in SQP 2.1. In the meta-analysis (Saris et al. 2011), it has been found that not only question characteristics, such as question wordings, response scales and labelling, but also the written and spoken language used in formulating the questions, determine the reliability and validity of questions.

With respect to nonWestern languages, some studies have evaluated different designs of response scales by the means of a split-ballot experiment (Lau 2016; Liao 2014). Some, such as Chen (2005) and Hsiao and Tu (2012), have applied MTMM to evaluate validity and reliability in Taiwan using Chinese-language content, but none had a focus on the effect of the formulation of single questions. Because no MTMM experiments have been done in Asia and specifically in Chinese, so far, the quality of survey questions cannot be predicted with SQP. Therefore, we designed this research to start by collecting quality estimates based on MTMM experiments in Taiwan. Previous studies have indicated that respondents in East Asian countries, for example, tend to more frequently choose responses in the middle of the scale than those in the West because of

the influence of collectivism (Chen et al. 1995; Harzing 2006). It is unknown whether the same quality of survey instruments can be achieved in both a different cultural context and writing system. The cultural transportability of experimental and pretesting techniques cannot be assumed; it has to be tested. For instance, Goerman and Caspar (2010) and Pan et al. (2005) have found that cognitive interviewing does not work equally well across cultures.

Using data from the SB-MTMM design in the Taiwan Social Change Survey (TSCS) and corresponding data from the 2002 ESS Round 1 in Denmark, this study aims to compare the measurement quality of different response designs across countries. It is of interest to explore the similarity, as well as differences, when the same experimental approach is applied. Therefore, we conduct a test for measurement invariance with the aim of concluding whether relationships and means across countries can be compared. The next section discusses the SB-MTMM approach that is used for this study and briefly introduces the test for measurement invariance. We then present the research design and results. Discussion on the findings are provided.

2. The SB-MTMM Design

A drawback of the classic MTMM design is the burden on each respondent of being asked multiple questions that assess the same construct. In addition, early questions may influence answers to later questions due to memory that is carried over. Consequently, the data quality may be overestimated (Saris et al. 2004). In order to minimize the carry-over effect from the previous answer, an interval of at least 20 minutes between the administration of the related items (Van Meurs and Saris 1990) is suggested.

The SB-MTMM design reduces the cognitive burden on respondents by using two, rather than three, methods in the MTMM design, while three traits are measured. Random samples of the same population are also used, as in the split-ballot experiments, but each respondent only needs to answer the questions concerning the same trait twice. This is seen to combine the benefits of the split-ballot approach and the MTMM approach, in that it enables researchers to evaluate measurement bias, reliability, and validity simultaneously, while reducing response burden (Saris and Gallhofer 2014).

We assume that the estimation model for the SB-MTMM design is the same as in the standard approach, given that the random samples are drawn from the same population. In the standard MTMM design, a minimum of three traits are measured using three different methods, leading to nine different observed variables. Therefore, a correlation matrix of 9×9 is obtained. However, this is not always the case when using the SB-MTMM approach. Nevertheless, the same models can be used, as we will show later. Various models have been suggested for analysis of the correlation matrices, and a true score model proposed by Saris and Andrews (1991) is commonly applied. The advantage of this model is that its standardization of the coefficients directly provides the estimates of the reliability and validity coefficients (Saris and Gallhofer 2014). Recent applications include those by Revilla and Saris (2013), Saris et al. (2010), Zavala-Rojas et al. (2018), Revilla (2015), Revilla et al. (2015), and Oberski et al. (2007).

The use of a minimum of three traits to be repeated using at least three methods serves the purpose of identification. With such a consideration, the model can be defined by the

following Equations (Saris and Andrews 1991):

$$Y_{ij} = r_{ij}T_{ij} + e_{ij} \tag{1}$$

$$T_{ij} = v_{ij}F_i + m_{ij}M_j \tag{2}$$

where Y_{ij} is the observed variable for the ith trait and the jth method; r_{ij} and v_{ij} are the reliability and validity coefficients for the ith trait and the jth method, respectively; T_{ij} is the true score or systematic component of the response Y_{ij}; e_{ij} is the random error associated with Y_{ij}; F_i is the ith trait (or factor); M_j is the variation in scores due to the jth method; and m_{ij} is the method effect for the ith trait and the jth method. The model posits that the observed variable is the sum of the systematic component plus a random error. Also, the systematic component of a response is the sum of the trait and the effect of the method used to assess it.

We make the assumption that the traits are correlated with one another. The random errors are not correlated with one another, nor with the independent variables in the different equations. The method factors are assumed to not be correlated with one another, nor with the traits or the random errors. Figure 1 is a graphical presentation of the true score model.

When all variables other than e_{ij} are standardized, v_{ij}, and m_{ij} correspond to the reliability, validity, and method effect coefficients, respectively, of a measure, while the squares of these coefficients present the reliability, validity and the method variance, respectively. In this approach, the reliability and validity of single questions have been evaluated, not complex concepts. As a result, the validity does not indicate how well the measured indicator represents the concept of interest. The validity is only affected by the method used, that is, $v_{ij}^2 = 1 - m_{ij}^2$. The lack of reliability will decrease the correlations between the variables, while the method effects will increase the correlations between the variables measured by the same or similar methods. This effect is called "common method variance". The model specified in Equation 1 and Equation 2 assumes that the disturbance term only contains a random error component, e_{ij}. Therefore, in this model, we make

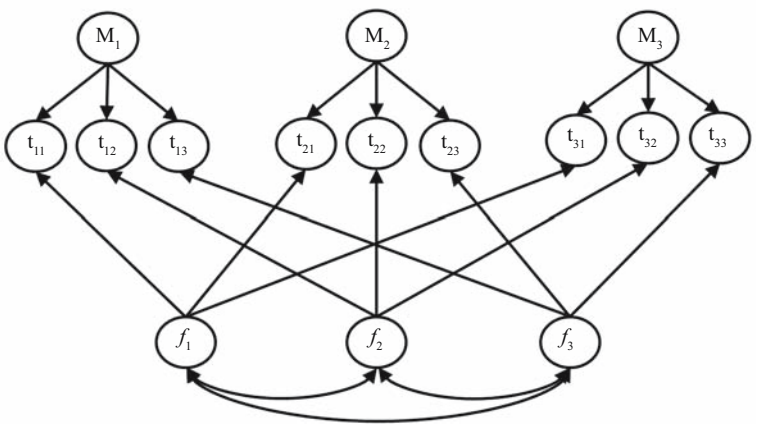

Fig. 1. MTMM model illustrating the true scores and their factor of interest.

the assumption that there is no unique component (Saris and Andrews 1991, 579). This assumption is plausible when the stem of the questions remains the same in the multitrait-multimethod experiment, and the variation only comes from the variations in the methods.

The total quality of a measure can then be computed as $q_{ij}^2 = r_{ij}^2 \times v_{ij}^2$, where q_{ij}^2 represents the amount of the variance of the observed variable, which is explained by the latent trait of interest. The quality indicators, reliability, and validity are typical measures of quality, which vary between zero and one, like correlation coefficients. With respect to the multiple groups in the SB-MTMM design, estimates for the parameters of the model can be obtained using structural equation modeling for multiple-group analysis (Saris and Gallhofer 2007, 2014).

3. Test for Measurement Invariance

With the SB-MTMM model, the variance-covariance matrix of the traits, F_i, is obtained. This correlation matrix is corrected for measurement error and can be used to test whether the same construct is measured across countries. The test for measurement invariance is typically done using the variance-covariance matrix of observed variables, although a criticism that has been referred to as *susceptibility,* that is, to what extent the procedure is sensitive to artifacts in the response process, is commonly made (Butts et al. 2006; Marsh and Byrne 1993; Byrne and Watkins 2003; Saris and Gallhofer 2014). Saris and Gallhofer (2014 chap. 16) showed that in a test for measurement invariance, the *response* process can be distinguished from the *cognitive* process. As we have said above, the variance-covariance matrix corrected for measurement errors will be obtained in the MTMM analysis, and this matrix can be used to test for the cognitive equivalence or comparability of the concepts in the different countries.

Therefore, we used the variance-covariance matrix of the latent traits to test for measurement invariance. The test is usually conducted in three steps, where each step is a prerequisite of the next one. In the first step, a *configural* model is fitted to check whether the pattern of fixed and free loadings and disturbance terms is the same across groups (Horn and McArdle 1992). In the second step, *metric invariance*, the configural model is restricted to one where the factor loadings of equivalent manifest variables are invariant across countries. When the model is not rejected, comparisons of relationships across groups can be made (Horn and McArdle 1992). The third step, *scalar invariance,* implies that, in addition to invariance in the factor loadings, intercepts of equivalent manifest variables are also restricted to be the same across groups. If the model is not rejected, comparisons of means can also be made across groups.

Figure 2 shows the path diagram of the model to test for measurement invariance. The model is specified in Equations (3) to (5).

$$f_1 = \tau_1 + \eta_1\lambda_1 + d_1 \tag{3}$$

$$f_2 = \tau_2 + \eta_1\lambda_2 + d_2 \tag{4}$$

$$f_3 = \tau_3 + \eta_1\lambda_3 + d_3 \tag{5}$$

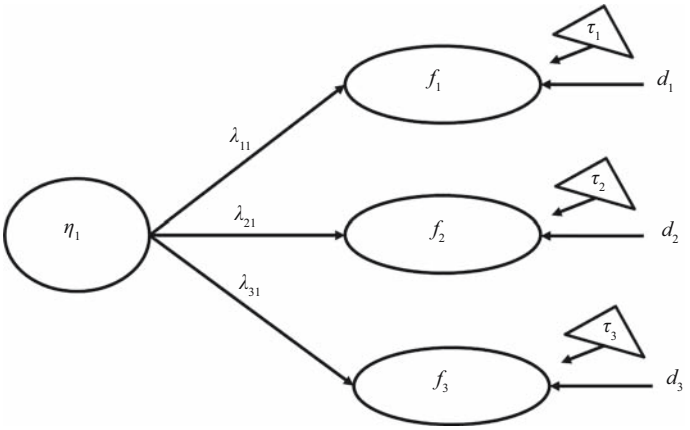

Fig. 2. Model to test for measurement invariance.

Where η_1 is the concept of interest and F1 to F3 represent the indicators used in the study corrected for measurement errors.

Standard restrictions were imposed to identify the model: the loading of the first trait (λ_1) and its corresponding intercept (d_1) were fixed to one and zero respectively. Secondly, we make the assumption that the error terms are not correlated with each other or with the latent variables. To test for metric invariance, we assume that the loadings (λ) are equal across groups, and for the scalar invariance we assume that the intercepts (d) are also equal across groups.

4. Research Design

Two data sources are used for this study, one from Taiwan and the other from Denmark, and both are collected using the computer-assisted personal interview (CAPI) technique. Taiwanese data are drawn from the 2015 Taiwan Social Change Survey (TSCS) (Fu et al. 2016), which included questions on globalization, work, family, and mental health, and included the SB-MTMM experiment. Surveys were delivered to randomly selected adults aged 18 years or older within each of the selected municipalities. A three-stage stratified sampling design was adopted based on the urbanization level and geographic areas of the townships and boroughs in Taiwan as the primary sampling unit (PSU). The probability proportional to size (PPS) sampling method was used in the first two stages – township and village or *li* under townships, respectively. Finally, household-registered residents in each village or *li*, which are equivalent-sized neighborhoods in urban areas, are systematically selected to obtain a representative sample of Taiwan's population. A total of 2,034 complete cases are obtained, with a response rate of 57%.

The experiment conducted in the 2015 TSCS adopted a two-group SB-MTMM design. The sample was randomly divided into two subsamples based on the respondent's number, which was assigned beforehand, as odd or even. The odd-numbered subsample (Sample 1) got Method 1 (M_1) first and then Method 3 (M_3), and the even-numbered subsample (Sample 2) got Method 2 (M_2) first, but Method 3 (M_3) next. As shown in Table 1, the combination of M_1 and M_2 was missing by design.

Table 1. Two-group SB-MTMM design.

	Method 1	Method 2	Method 3
Method 1	Sample 1		
Method 2	NONE	Sample 2	
Method 3	Sample 1	Sample 2	Sample 1 + 2

In other words, this set of correlations between the variables measured by M_1 and the variables measured by M_2 is absent. Saris et al. (2004) have shown, based on the work of Satorra (1993), that, in general, the parameters of this model evaluated with two groups are identified and all quality indicators can be estimated using multiple group estimation, except when the correlations between the traits are very similar or zero (Revilla and Saris 2013).

The measures for the SB-MTMM experiment include several questions. Each question is measured with two sets of response scales (M_1 and M_2 in the case of 2015 TSCS) that are answered by Sample 1 and Sample 2, respectively, and one other set (M_3) answered by all of the respondents. Both of the subsamples answer all other questions in the survey as well.

The questions used are commonly used indicators of the latent concept "Political satisfaction". The following three indicators of political satisfaction are used for the experimental design as follows:

1. On the whole, how satisfied are you with the present state of the economy in [country]?
2. Now thinking about the [country] government, how satisfied are you with the way it is doing its job?
3. And, on the whole, how satisfied are you with the way democracy works in [country]?

Using the same indicators for the three methods, M_1 is measured using a fully labeled four-point scale, with labels very satisfied, satisfied, dissatisfied and very dissatisfied. M_2 and M_3 are measured from 0 to 5 and from 0 to 10, respectively, both using show cards with only the endpoints labeled as "extremely dissatisfied" and "extremely satisfied". For all of the methods, a higher score indicates a higher level of satisfaction. The correlation matrices were obtained for analysis using a structural equation model. The design of the experiment has been summarized in Figure 3, where Ts_i denotes the Taiwan sample i, Ds_i denotes the Danish sample i where $i = 1, 2$, and c stands for the combination of the two samples within each country.

In order to estimate the parameters, covariance matrices obtained for the nine measures are used in the multi-group SEM in LISREL. The maximum likelihood (ML) approach is adopted to deal with missing data, which occurs by design (Saris et al. 2004).

The same measures of satisfaction, use of show cards and a two-group SB-MTMM experimental design can be found in the 2002 ESS Round 1. The data from Denmark (ESS1_DK) are used for the comparison with the data in Taiwan due to the same data collection mode of CAPI in both the main questionnaire and supplemental questions. The

Method / Question	M_1 (4-point scale) (01) Very dissatisfied (02) Fairly dissatisfied (03) Fairly satisfied (04) Very satisfied	M_2 (6-point scale) (00) Extremely dissatisfied ⋮ (05) Extremely satisfied	M_3 (11-point scale) (00) Extremely dissatisfied ⋮ (10) Extremely satisfied
Q1. How satisfied with present state of economy in country	TS1/DSC	TS2/DS1	TSC/DS2
Q2. How satisfied with the national government			
Q3. How satisfied with the way democracy works in country			

Fig. 3. Two-group SB-MTMM design for Denmark and Taiwan.

sampling design for ESS1_DK is a simple random sample based on a register-based sampling frame, with a lower age cut-off of 16 years. A total of 1,506 complete cases were obtained, with a response rate of 67.56%. More details can be found in the ESS1 – Documentation Report (European Social Survey 2014, 42–47). As in the case of Taiwan, the SB-MTMM experiment was performed alongside other questions, among others, about politics, work, family, well-being and immigration.

One difference in the experimental design between 2015 TSCS and ESS1_DK is that all the respondents in the latter got M_1 first, and then M_2 and M_3 for samples 1 and 2, respectively. Therefore, the combination of M_2 and M_3 is missing by design in the Danish data. The differences in the data structures are clearly observable in the correlation matrices presented in Table 3 and Table 4. Nevertheless, the same model can be estimated on the basis of these two different correlation matrices.

Another difference is that in Taiwan, an unfolding technique, in which interviewers first asked about the direction and then about the degree of attitudes (Schaeffer and Presser 2003), was used for M_1, with the scale coded reversely as 1 = very satisfied to 4 = very dissatisfied. In the ESS, one direct question was used, in which all four categories were presented immediately. Although the data have been recoded to have the same response order as that in the Danish data, the difference in procedure means that we were not able to determine the effect of the scale length, only because this effect is confounded with other aspects. However, it is possible to determine which measure is better in each country and across countries.

5. Results of the SB-MTMM Experiment

Socio-demographic variables were first compared between Denmark and Taiwan with post-stratified weights. As shown in Table 2, the distributions of demographic characteristics are similar in age and gender. The proportions of those aged 60 years or older are lower when compared to other age groups. Also, there are similar proportions of men and women in both samples. On the other hand, the proportions of married and widowed respondents in Taiwan are higher, but those of single or divorced respondents are higher in Denmark. It is noted that ESS used other variables to ask respondents whether they live with a partner, but a category of "cohabitant" is included in TSCS for marital

Table 2. Description of Denmark and Taiwan samples.

Country	Denmark		Taiwan	
Variable	*f*/M	%/SD	*f*/M	%/SD
Age				
16–29 years	304	20.2%	401	19.8%
30–39 years	277	18.4%	359	17.7%
40–49 years	245	16.3%	382	18.8%
50–59 years	294	19.5%	378	18.6%
60–79 years	197	13.1%	263	13.0%
70 years or older	189	12.5%	247	12.2%
P = .553	N = 1506	100%	N = 2030	100%
Gender				
Female	736	49%	1043	51.3%
Male	766	51%	991	48.7%
P = .096	N = 1502	100%	N = 2034	100%
Marital***				
Single, never married	486	31.3%	594	29.3%
Married	800	53.6%	1141	56.2%
Divorced	119	8.0%	123	6.1%
Separate	15	1.0%	9	0.4%
Widowed	91	6.1%	149	7.3%
Cohabitant	0	0%	14	0.7%
P = .000	N = 1493	100%	N = 2030	100%
Educational level***				
Elementary or less	26	1.7%	416	20.5%
Junior high schoold	392	26.2%	256	12.6%
Senior high school	733	49%	541	26.7%
Tertiary education or higher	344	23%	815	40.2%
P = .000	N = 1495	100%	N = 2028	100%
Health status***				
Very good	648	43.2%	468	23.0%
Good	512	34.2%	619	30.5%
Fair	253	16.9%	630	31.0%
Bad	86	5.7%	314	15.5%
P = .000	N = 1499	100%	N = 2031	100%
Interested in Politics***				
Very interested	202	13.4%	34	1.7%
Quite interested	723	48.1%	330	16.3%
Hardly interested	487	32.4%	654	32.4%
Not at all interested	90	6.0%	1003	49.6%
P = .000	n = 1502	100%	n = 2021	100%

Table 2. Continued.

Country	Denmark		Taiwan	
Variable	*f*/M	%/SD	*f*/M	%/SD
Method 1 (4-point scale)				
Q1. How satisfied with present state of economy in country***	2.91	0.577	2.12	0.771
Q2. How satisfied with the national government***	2.72	0.656	2.04	0.767
Q3. How satisfied with the way democracy works in country***	3.10	0.595	2.58	0.752
Method 2 (6-point scale)				
Q1. How satisfied with present state of economy in country***	3.46	0.917	1.96	1.241
Q2. How satisfied with the national government***	3.06	1.091	1.82	1.254
Q3. How satisfied with the way democracy works in country***	3.74	0.854	2.77	1.339
Method 3 (11-point scale)				
Q1. How satisfied with present state of economy in country***	6.92	1.938	3.94	2.036
Q2. How satisfied with the national government***	5.86	2.268	3.58	2.089
Q3. How satisfied with the way democracy works in country***	7.24	1.876	5.31	2.251

*p < .05.
**p < .01.
***p < .001.

status. If "cohabitant" is dropped, the proportions of other categories in marital status increase slightly, from 0.06% to 0.4%, in TSCS and the result of the Chi-square test remains significant. Educational levels are recoded for both ESS1_DK and TSCS for comparison. Almost half of the Danish sample have a senior high school degree (49%), while 40% of the respondents in the Taiwanese data have a higher tertiary education degree, including formal education at colleges, universities and higher degrees.

There are other variables that can be used to reveal the difference between Denmark and Taiwan, such as health status and interest in politics. Both ESS1_DK and 2015 TSCS employed five response categories for health status, but the former used a balanced scale, from "very good" to "very bad" with "fair" as the middle response, while the latter used an unbalanced one, from "excellent" to "bad." The categories of "bad" and "very bad" in ESS1_DK are combined and so are "excellent" and "very good" in 2015 TSCS, resulting in four response categories (see Table 2).

With regard to interest in politics, both samples employed the same response categories for measurement. It is noticeable that more than 60% of the Danish sample indicated certain levels of interest in politics, while nearly half of the Taiwanese sample were not at all interested in politics.

Table 3. Correlations, means, and standard deviations of Danish samples[1].

Sample 1	M_1			M_2			M_3		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
M_1									
Q1	1								
Q2	.410	1							
Q3	.288	.288	1						
M_2									
Q1	.414	.267	.179	1					
Q2	.262	.677	.162	.410	1				
Q3	.269	.261	.473	.407	.388	1			
M_3									
Q1	.0	.0	.0	.0	.0	.0	1		
Q2	.0	.0	.0	.0	.0	.0	.0	1	
Q3	.0	.0	.0	.0	.0	.0	.0	.0	1
Mean	2.91	2.73	3.11	6.94	5.92	7.31	.0	.0	.0
S.D.	.579	.652	.599	1.915	2.273	1.884	1.0	1.0	1.0
n	653	653	653	653	653	653			
Sample 2	M_1			M_2			M_3		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
M_1									
Q1	1								
Q2	.497	1							
Q3	.411	.317	1						
M_2									
Q1	.0	.0	.0	1					
Q2	.0	.0	.0	.0	1				
Q3	.0	.0	.0	.0	.0	1			
M_3									
Q1	.554	.373	.300	.0	.0	.0	1		
Q2	.388	.744	.196	.0	.0	.0	.466	1	
Q3	.362	.309	.603	.0	.0	.0	.421	.372	1
Mean	3.46	3.08	3.74	.0	.0	.0	6.92	5.95	7.30
S.D.	.911	1.069	.844	1.0	1.0	1.0	1.962	2.203	1.794
n	687	687	687				687	687	687

[1]All of the correlation coefficients are significant at the .000 level.

Table 4. Correlations, means, and standard deviations of Taiwanese samples[1].

Sample 1	M_1			M_2			M_3		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
M_1									
Q1	1								
Q2	.678	1							
Q3	.317	.388	1						
M_2									
Q1	.0	.0	.0	1					
Q2	.0	.0	.0	.0	1				
Q3	.0	.0	.0	.0	.0	1			
M_3									
Q1	.525	.518	.304	.0	.0	.0	1		
Q2	.496	.628	.360	.0	.0	.0	.733	1	
Q3	.252	.314	.532	.0	.0	.0	.428	.482	1
Mean	2.09	2.02	2.57	.0	.0	.0	4.07	3.68	5.41
S.D.	.749	.755	.757	1.0	1.0	1.0	1.963	2.039	2.219
n	880	880	880				880	880	880
Sample 2	M_1			M_2			M_3		
	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
M_1									
Q1	.0	.0	.0						
Q2	.0	.0	.0						
Q3	.0	.0	.0						
M_2									
Q1	.0	.0	.0	1					
Q2	.0	.0	.0	.686	1				
Q3	.0	.0	.0	.381	.475	1			
M_3									
Q1	.0	.0	.0	.673	.637	.345	1		
Q2	.0	.0	.0	.589	.748	.399	.781	1	
Q3	.0	.0	.0	.293	.337	.651	.392	.419	1
Mean	.0	.0	.0	1.97	1.80	2.77	3.77	3.44	5.21
S.D.	1.0	1.0	1.0	1.216	1.239	1.331	2.067	2.100	2.298
n				890	890	890	890	890	890

[1]All of the correlation coefficients are significant at the .000 level.

As for the satisfaction measured by three response scales, significant differences are found between countries, as well as among methods. Among different methods, the mean scores of three satisfaction questions are higher in Denmark than in Taiwan. In particular, the differences between Denmark and Taiwan are larger when satisfaction is measured by M_2 and M_3, despite the consistently low levels of satisfaction in the Taiwanese sample.

The results of the satisfaction measures using a two-group SB-MTMM design are reported in Table 3 and Table 4 for ESS1_DK and 2015 TSCS, respectively, indicating incomplete data in each of the subsamples. Since both of the datasets employed approximately the same SB-MTMM experimental design, the parameter estimation followed the same procedure. The correlations for the unobserved variables are indicated by zeros and the variances by ones, as required for the multiple-group analysis with incomplete data in LISREL (Allison 1987). The correlation between the variables measured by M_1 and M_2 is missing by design. Therefore, the parameters are estimated based on the incomplete covariance matrix. In addition, in order to estimate the coefficients of reliability, validity, and method effects for the two randomly selected subsamples simultaneously, we make the assumption that the model is the same for both groups, except for the specification of selecting the variables of the two groups.

The Taiwanese data had the peculiarity that the correlations between questions 1 and 2 were much higher than the correlation between these variables and question 3. The program Jrule (Van der Veld et al. 2008), which was used to detect misspecifications in the model, also detected this high correlation and suggested introduction of a correlated error between questions 1 and 2 in the model for the Taiwanese data. Only in Method 2 was this correlated error not significantly different from zero. In the other two methods, these correlations were 0.14 for M_1 and 0.28 for M_3. The explanation is not so simple, but it is clear that a deteriorating economy has been a serious issue in Taiwan in the past decade, and this has been seen as the responsibility of the government. Research in political science has indicated such consequences of economic performance on voting behavior and named it "economic voting" (Wu and Lin 2013). One can therefore expect a much higher correlation between satisfaction with the government and satisfaction with the economy than between these two and the functioning of democracy, which does not depend so much on the present government only. With this one correction, a proper solution is obtained with a $chi^2 = 32.20$ and $df = 38$ after we corrected for the zero cells in the correlation matrices, the RMR = .011. The Jrule approach to test for local misspecifications (Saris et al. 2009) did not suggest improvements.

Table 5. Estimates of the parameters for the two-group SB-MTMM design[1].

	Reliability						Validity					
Method	M_1		M_2		M_3		M_1		M_2		M_3	
Country	D	T	D	T	D	T	D	T	D	T	D	T
Q1	.55	.55	.74	.72	.58	.72	.79	.86	.92	.98	.74	.81
Q2	.74	.69	.90	.87	.77	.79	.87	.90	.94	.98	.81	.83
Q3	.91	.62	.62	.85	.76	.62	.76	.88	.90	.98	.81	.81
Average	.73	.62	.75	.81	.70	.71	.81	.88	.92	.98	.79	.82

[1]"D" denotes the ESS1_DK data and "T" denotes the 2015 TSCS.

Table 6. The quality of the different questions for the different methods used[1].

	M_1		M_2		M_3	
Country	D	T	D	T	D	T
Q1	0.43	0.47	0.68	0.71	0.45	0.58
Q2	0.64	0.62	0.85	0.85	0.62	0.66
Q3	0.69	0.54	0.55	0.83	0.62	0.50
Average	0.59	0.54	0.69	0.8	0.56	0.58

[1]"D" denotes the ESS1_DK data and "T" denotes the 2015 TSCS.

The estimated reliabilities and validities for ESS1_DK and 2015 TSCS are reported in Table 5. As the total quality of a measure is defined as follows:

$q_{ij}^2 = r_{ij}^2 \times v_{ij}^2$, Table 6 has been derived from Table 5.

Table 6 shows that in both countries, the second method, the six-point scale has better quality on average over the three questions. This is also true for three questions, except for question 3, in Denmark. The quality of the measures using Method 1 (four-point scale) and Method 3 (eleven-point scale) do not differ very much. However, when we look at the estimated validities of the questions using Method 1 and Method 3 we see that they are slightly better in Taiwan, while the reliability using Method 1 in Taiwan is lower.

5.1. *Results of the Test for Measurement Invariance*

Table 7 shows the variance-covariance matrix of the latent traits. As the baseline model has only recently been identified, it is not possible to conduct a robustness test with the configural model. When the loadings are restricted, Jrule shows that the loading of the second trait is misspecified. As was mentioned above, the government is seen as responsible for the economic situation. This seems to be stronger for the case in Taiwan

Table 7. The variance-covariance matrix of the traits.

	Denmark		
n = 1502	Q_1	Q_2	Q_3
Q_1	.17		
Q_2	.13	.34	
Q_3	.09	.09	.19
Mean	2.91	2.73	3.11
	Taiwan		
n = 2020	Q_1	Q_2	Q_3
Q_1	.28		
Q_2	.28	.37	
Q_3	.15	.19	.29
Mean	2.09	2.02	2.57

Table 8. Likelihood ratio test of the metric and scalar models.

	Chi-square	Chi-square difference	DF difference	Pr. (>Chi^2)
Partially metric invariant model	0.0766			
Partially scalar invariant model	0.8217	0.74511	1	0.388

than in Denmark. This means that this measurement instrument has only partial metric invariance, that is, only the first and the third items are comparable across countries. Leaving this loading free, the concepts can be seen as comparable across the two countries.

As equality of loadings is a prerequisite for scalar invariance, the intercepts are restricted to being equal in both countries, except in the second trait. The likelihood ratio test (Table 8) indicates that the fit of the scalar model is not significantly different from the one of the metric model, and Jrule did not show additional misspecifications. These results imply that, at the cognitive level, partial scalar invariance is established. As the observed data are corrected for measurement errors, this result means that the relationships between these concepts and other variables and the latent means of the concepts can be compared across countries.

6. Conclusion

The SB-MTMM approach has been widely applied in the ESS to evaluate the quality of survey measures. It remains unclear whether this approach performs equally well in logographic writing systems. In addition, it is of interest to explore possible similarity or difference between ESS and Taiwanese data. Using a two-group experimental design, Danish data from ESS Round 1 and the 2015 TSCS are compared. The results indicated that questions measured by six-point scales with labels at endpoints (M_2) have the best quality, while the measures on either a four-point scale with full labels or an eleven-point scale are equally acceptable. Although differences between Danish and Taiwanese data can be observed, the findings are comparable, despite the fact that the order of applying methods differed. These findings are contrary to previous research suggesting that fully labeled response scales provide higher reliability than those with endpoints (Alwin and Krosnick 1991; Holbrook et al. 2006; Weng 2004), while the results are consistent with other studies (Saris and Gallhofer 2014; Saris et al. 2004). The designs of response scale deserve further examination. However, these methods differed in more than just this one aspect, which may explain these results.

One possible reason for the relatively poor quality of M_1 in the 2015 TSCS may be the different measurement procedures used for the different methods during the face-to-face interview. An unfolding technique, in which interviewers first asked about the direction and then about the degree of attitudes (Schaeffer and Presser 2003), is used for M_1 to minimize the tendency of choosing the middle category. On the other hand, show cards are provided upon request for M_2 and M_3, so it is easier for the respondents to answer the questions using Methods 2 and 3, rather than using Method 1. While the inquiry process

should be considered as part of the methods, researchers need to be cautious with its influence on data quality.

One cannot draw general conclusions about the effect of different aspects of the methods on the quality of questions, because often, like here, more aspects vary at the same time. Also, findings on quality of measures may differ by the measured topics. For general conclusions, we refer to the results of meta-analyses over large numbers of MTMM experiments (Saris and Gallhofer 2014).

A second result is that the concept "political satisfaction" is only partially invariant across the two countries. The results of the invariance test show that the understanding of the indicators of satisfaction with the economy and with the way democracy works are comparable in Denmark and Taiwan. However, this is not the case for satisfaction with the government. For this last indicator, there seems to be a different interpretation in the two countries. This signifies that means and relationships of the latent variable "political satisfaction" can be compared across countries, but composite scores can only be compared if one uses only the comparable indicators in computing the composite scores.

7. References

Allison, P.D. 1987. "Estimation of Linear Models with Incomplete Data." In *Sociological Methodology*, edited by C.C. Clogg, 71–103. Washington DC: American Sociological Association. Doi: https://doi.org/10.2307/271029.

Alwin, D.F. 1974. "Approaches to the Interpretation of Relationships in the Multitrait Multimethod Matrix." *Sociological Methodology* 5: 79–105. Doi: https://doi.org/10.2307/270833.

Alwin, D.F. 1997. "Feeling Thermometers versus 7-point Scales: Which Are Better?" *Sociological Methods and Research* 25: 318–340. Doi: https://doi.org/10.1177/0049124197025003003.

Alwin, D.F. and J.A. Krosnick. 1991. "The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes." *Sociological Methods and Research* 20: 139–181. Doi: https://doi.org/10.1177/0049124191020001005.

Bjørnskov, C. 2010. "How Comparable Are The Gallup World Poll Life Satisfaction Data?" *Journal of Happiness Studies* 11: 41–60. Doi: https://doi.org/10.1007/s10902-008-9121-6.

Butts, M.M., R.J. Vandenberg, and L.J. Williams. 2006. "Investigating the Susceptibility of Measurement Invariance Tests: The Effects of Common Method Variance." *Academy of Management Proceedings* 2006(1): D1–D6. Doi: https://doi.org/10.5465/AMBPp.2006.27182126.

Byrne, B.M. and D. Watkins. 2003. "The Issue of Measurement Invariance Revisited." *Journal of Cross-Cultural Psychology* 34(2): 155–175. Doi: https://doi.org/10.1177/0022022102250225.

Campbell, D.T. and D.W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56(2): 81–105. Doi: https://doi.org/10.1037/h0046016.

Chen, C.K. 2005. "Construct Model of Knowledge: Based Economy Indicators." *Management Review* 24(3): 17–41. Doi: https://doi.org/10.6656/MR.2005.24.3.CHI.17.

Chen, C., S.Y. Lee, and H.W. Stevenson. 1995. "Response Style and Cross-Cultural Comparisons of Rating Scales among East Asian and North American Students." *Psychological Science* 6: 170–175. Doi: https://doi.org/10.1111/j.1467-9280.1995.tb00327.x.

ESS Round 1: European Social Survey. 2014. *ESS-1 2002 Documentation Report*. Edition 6.4. Bergen, European Social Survey Data Archive, NSD – Norwegian Centre for Research Data for ESS ERIC. Available at: http://www.europeansocialsurvey.org/docs/round1/survey/ESS1_data_documentation_report_e06_4.pdf (accessed May 2016).

Fu, Y.-C., Y.-H. Chang, S.-H. Tu, and P.-S. Liao. 2016. *2015 Taiwan Social Change Survey (Round 7, Year 1): Globalization, Work, Family, Mental Health, and Political Participation* (C00315_2) [Data file]. Available at Survey Research Data Archive, Academia Sinica. Doi: https://doi.org/10.6141/TW-SRDA-C00315_1-1.

Goerman, P.L. and R.A. Caspar. 2010. "Managing the Cognitive Pretesting of Multilingual Survey Instruments: A Case Study of Pretesting of the U.S. Census Bureau Bilingual Spanish/English Questionnaire." In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. Harkness, et al.: 75–90. John Wiley and Sons, Inc. Doi: https://doi.org/10.1002/9780470609927.ch5.

Harzing, A.W. 2006. "Response Styles in Cross-National Survey Research: A 26 Country Study." *International Journal of Cross Cultural Management* 6 (2)(August 1): 243–266. Doi: https://doi.org/10.1177/1470595806066332.

Hsiao, C.-C. and C.-H. Tu. 2012. "Common Method Variance in the Measurement of Teachers' Creative Teaching." *Psychological Testing* 59(4): 609–639. Doi: http://dx.doi.org/10.7108%2fPT.201212.0609.

Holbrook, A., Y.K. Cho, and T. Johnson. 2006. "The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties." *Public Opinion Quarterly* 70: 565–595. Doi: https://doi.org/10.1093/poq/nfl027.

Horn, J.L. and J.J. McArdle. 1992. "A Practical and Theoretical Guide to Measurement Invariance in Aging Research." *Experimental Aging Research* 18(3–4): 117–144. Doi: https://doi.org/10.1080/03610739208253916.

Lau, C.Q. 2016. "Rating Scale Design among Ethiopian Entrepreneurs: A Split-Ballot Experiment." *International Journal of Public Opinion Research* edw031. Doi: https://doi.org/10.1093/ijpor/edw031.

Liao, P.-S. 2014. "More Happy or Less Unhappy? Comparison of the Balanced and Unbalanced Designs for the Response Scale of General Happiness." *Journal of Happiness Studies* 15(6): 1407–1423. Doi: https://doi.org/10.1007/s10902-013-9484-1.

Marsh, H.W. and B.M. Byrne. 1993. "Confirmatory Factor Analysis of Multitrait-Multimethod Self-concept Data: Between-group and Within-group Invariance Constraints." *Multivariate Behavior Research* 28(3): 313–449. Doi: https://doi.org/10.1207/s15327906mbr2803_2.

Van Meurs, A. and W.E. Saris. 1990. "Memory Effects in MTMM Studies." In *Evaluations of Measurement Instruments by Metaanalysis of Multitrait-Multimethod Studies*, edited by W.E. Saris and A. van Meurs, 134–146. Amsterdam: North Holland.

Oberski, D.L., W.E. Saris, and J. Hagenaars. 2007. "Why Are There Differences in Measurement Quality across Countries?" In *Measuring Meaningful Data in Social Research*, edited by G. Loosveldt and Swyngedouw. Leuven: Acco. Available at:

http://daob.nl/wp-content/uploads/2013/03/Oberski-Saris-Why-are-there-differences-in-measurement-quality-across-countries.pdf (accessed January 2019).

Oberski, D., W.E. Saris, and J.A. Hagenaars. 2010. "Categorization Errors and Differences in the Quality of Questions in Comparative Surveys." In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. Harkness, et al.: 435–453. Hoboken, NJ: Wiley. Doi: https://doi.org/10.1002/9780470609927.ch23.

Pan, Y., B. Craig, and S. Scollon. 2005. "Results from Chinese Cognitive Interviews on the Census 2000 Long Form: Language, Literacy, and Cultural Issues." *Statistical Research Division's Research Report Series* (Survey Methodology 2005 – 09). Washington, DC: U.S. Bureau of the Census. Available at https://www.census.gov/srd/papers/pdf/rsm2005-09.pdf (accessed November 2017).

Petersen, T. 2008. "Spilt Ballot as An Experimental Approach to Public Opinion Research." In *The Sage Handbook of Public Opinion Research*, edited by W. Donsbach and M.W. Traugott, 322–329. Los Angeles, CA: Sage. Available at: http://methods.sagepub.com/book/sage-hdbk-public-opinion-research/n30.xml (accessed January 2019).

Revilla, M. 2015. "Comparison of the Quality Estimates in a Mixed-Mode and a Unimode Design: An Experiment from the European Social Survey." *Quality and Quantity* 49(3): 1219–1238. Doi: https://doi.org/10.1007/s11135-014-0044-5.

Revilla, M. and W.E. Saris. 2013. "The Split-Ballot Multitrait-Multimethod Approach: Implementation and Problems." *Structural Equation Modeling: A Multidisciplinary Journal* 20: 27–46. Doi: https://doi.org/10.1080/10705511.2013.742379.

Revilla, M., W.E. Saris, G. Loewe, and C. Ochoa. 2015. "Can a Non-Probabilistic Online Panel Get Similar Question Quality as the ESS?" *International Journal of Market Research* 57(3): 395–412. Available at: https://www.mrs.org.uk/ijmr_article/article/104501 (accessed January 2019).

Saris, W.E. and F.M. Andrews. 1991. "Evaluation of Measurement Instruments Using a Structural Modeling Approach." In *Measurement Errors in Surveys*, edited by P.P. Biemer, et al.: 575–597. New York, NY: Wiley.

Saris, W.E. and I.N. Gallhofer. 2007. "Estimation of the Effects of Measurement Characteristics on the Quality of Survey Questions." *Survey Research Methods* 1: 29–43. Doi: http://dx.doi.org/10.18148/srm/2007.v1i1.49.

Saris, W.E. and I.N. Gallhofer. 2014. *Design, Evaluation, and Analysis of Questionnaires for Survey Research* (Second edition). Hoboken, NJ: Wiley.

Saris, W.E., A. Satorra, and G. Coenders. 2004. "A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design." *Sociological Methodology* 34: 311–347. Doi: https://doi.org/10.1111/j.0081-1750.2004.00155.x.

Saris, W.E., A. Satorra, and W.M. van der Veld. 2009. "Testing Structural Equation Models or Detection of Misspecifications?" *Structural Equation Modeling: A Multidisciplinary Journal* 16(4): 561–582. Doi: https://doi.org/10.1080/10705510903203433.

Saris, W.E., R. Veenhoven, A.C. Scherpenzeel, and B. Brunting. 2008. *A Comparative Study of Satisfaction with Life in Europe*. Budapest: Eötvös University Press.

Saris, W.E., M. Revilla, J.A. Krosnick, and E.M. Shaffer. 2010. "Comparing Questions with Agree/Disagree Response Options to Questions with Item-Specific Response

Options." *Survey Research Methods* 4: 61–79. Doi: https://doi.org/10.18148/srm/2010.v4i1.2682.

Saris, W., D. Oberski, M. Revilla, D. Zavala, L. Lilleoja, I. Gallhofer, and T. Gruner. 2011. "The Development of the Program SQP 2.0 for the Prediction of the Quality of Survey Questions." RECSM Working Paper 24, Universitat Pompeu Fabra. Available at: https://www.upf.edu/documents/3966940/3986764/RECSM_wp024.pdf (accessed January 2019).

Satorra, A. 1993. "Asymptotic Robust Inferences in Multi-sample Analysis of Augmented Moment Matrices." In *Multivariate Analysis; Future Directions*, edited by R. Rao and C.M. Cuadras, 211–229. Amsterdam, North Holland.

Schaeffer, N.C. and S. Presser. 2003. "The Science of Asking Questions." *Annual Review of Sociology* 29: 65–88. Doi: https://doi.org/10.1146/annurev.soc.29.110702.110112.

Schuman, H. and S. Presser. 1981. *Questions and Answers in Attitude Surveys*. New York: Academic Press.

Van der Veld, W., W.E. Saris, and A. Satorra. 2008. *Jrule 2.0: User Manual*, Unpublished document.

Weng, L-J. 2004. "Impact of the Number of Response Categories and Anchor Labels on Coefficient Alpha and Test-Retest Reliability." *Educational and Psychological Measurement* 64: 956–972. Doi: https://doi.org/10.1177/0013164404268674.

Wu, C-E. and Y-T. Lin. 2013. "Cross-Strait Economic Openness, Identity, and Vote Choice: An Analysis of the 2008 and 2012 Presidential Elections." *Journal of Electoral Studies* 20(2): 1–36. Doi: https://doi.org/10.6612/tjes.2013.20.02.01-35.

Zavala-Rojas, D., R. Tormos, W. Weber, and M. Revilla. 2018. "Designing Response Scales with Multi-Trait-Multi-Method Experiments." *Mathematical Population Studies* 25(2): 66–81. Doi: https://doi.org/10.1080/08898480.2018.1439241.

Received October 2016
Revised July 2018
Accepted July 2018