Two approaches to evaluate measurement quality in online surveys: An application using the Norwegian Citizen Panel

Anna DeCastellarnau

Melanie Revilla

RECSM Working Paper Number 53

May 2017

Two approaches to evaluate measurement quality in online surveys: An application using the Norwegian Citizen Panel

Anna DeCastellarnau, RECSM-Universitat Pompeu Fabra / Tilburg University

Melanie Revilla, RECSM-Universitat Pompeu Fabra

Abstract:
Previous research has shown that measurement errors are large in data collected through surveys. However, the size of these errors varies depending on the language, the mode of data collection, the question's characteristics, etc. Information about the size of these errors, or their complement, the measurement quality, can be used for designing better questionnaires, and to correct for measurement errors in substantive analyses. Getting further information about the measurement quality is therefore crucial.

However, little is known about the measurement quality of questions in web surveys. Thus, in this paper, we evaluate the measurement quality of a set of survey questions from the Norwegian Citizen Panel[1]; one of the few probability-based online panels existing at this day. To do so, we use two different approaches: (1) we implement two Multitrait-Multimethod experiments in this panel's 5th wave, and (2) we predict the quality of the questions using the software Survey Quality Predictor.

Overall, measurements' quality is quite high (between 0.60 and 0.89), even if comparing the results of the two approaches, we observe some minor differences in quality. In addition, using 11-point scales with two fixed reference points, usually, provides the highest quality.

Keywords: measurement quality, Multitrait-Multimethod (MTMM), SQP, web surveys, Norwegian Citizen Panel

Acknowledgements: We are thankful to the Norwegian Citizen Panel to accept our experimental proposal and specially to Sveinung Arnesen for his help in setting up the experiment and his valuable comments. The authors also appreciate the feedback from Willem Saris, Diana Zavala-Rojas and Teresa Queralt.

[1] The data applied in the analysis in this publication are based on "Norwegian Citizen Panel Wave 5, 2015". The survey was financed by the University of Bergen (UiB) and Uni Rokkan Centre. The data are provided by UiB, and prepared and made available by the Norwegian Social Science Data Services (NSD). Neither UiB, Uni Rokkan Centre nor NSD are responsible for the analyses/interpretation of the data presented here.

1. Introduction

Each researcher designing a survey should take a lot of different decisions like the mode of data collection, the exact formulation of the questions and their response scales. For instance, researchers need to determine how many answer categories to propose. The theory of information (Garner, 1960) states that in the case of bipolar concepts, a 2-point scale allows only the assessment of the direction of the attitude (e.g. satisfied versus dissatisfied), whereas a 3-point scale with a middle category allows the assessment of both the direction and the neutrality, and even more categories allow the assessment of its intensity or extremity (e.g. degrees of satisfaction or dissatisfaction). However, one can wonder till when the amount of information increases with the number of response categories. Krosnick and Fabrigar (1997) argued that even if too few categories fail to discriminate between respondents with different underlying opinions, too many categories may reduce the clarity of the meaning of the response options. Consequently, respondents tend to use only some response categories: for instance, on a scale from 0 to 100, most respondents will answer 30 or 75, and not 31 or 77. Overall, there is no agreement on how many answer categories to provide. Similarly, there is no agreement on the kind of labels that should be provided, the use of instructions, etc.

However, each choice is important because it can affect the respondents' answers. If researchers do not account for these effects in their data, the relationships between variables and the substantive conclusions will be biased. For instance, Saris and Gallhofer (2007) showed that the same questions, asked in the same country, in the same survey and to the same people, lead to opposite conclusions just because the number of response categories changed (p. 174). These differences can be explained by the different size of measurement errors when using different scales.

Borgatta and Bohrnstedt (1980) define measurement errors as a 'function of the fit between the manifest scale and the latent construct' (p. 153). Two types can be distinguished: random measurement errors, due to unintended and unpredicted mistakes of the respondents, interviewers or coders; and systematic measurement errors, due to the reaction of respondents to the variation of the method used (also called method effect).

Because surveys are commonly affected by both types of measurement errors, it is crucial for any survey to have information about their size (Saris & Gallhofer, 2014). First, this information is useful to develop better survey questions (Revilla, Zavala-Rojas, & Saris, 2016). However, even if the best possible survey questions were developed based on this knowledge, there will still be some errors. Thus, it is also necessary to correct for measurement errors in order to avoid misleading conclusions in substantive research (Saris & Revilla, 2016). This correction can be done in a simple way, as long as we first have information about the size of the measurement errors for the questions of interest (DeCastellarnau & Saris, 2014).

Instead of estimating directly the size of random and systematic measurement errors, we can also estimate the size of their complements: the measurement's reliability and validity, whose product is the measurement quality, also known as construct validity. Measurement quality is defined as the strength of the relationship between the latent variable of interest (e.g. satisfaction with democracy) and the observed answers to the survey question asked to measure this latent concept (e.g. How satisfied are you with the way the democracy works in your country? 1-Very satisfied, 2-Satisfied, 3-Dissatisfied, 4-Very dissatisfied). Said

differently, measurement quality is the proportion of explained variance due to the latent concept of interest. The observed variable will only measure perfectly the latent variable of interest, when both reliability and validity are one, i.e. when random and systematic errors are zero. This is very unlikely. In fact, Andrews (1984) found that 'about two-thirds of the survey measures examined contained between 50 percent and 83 percent valid variance' (p. 425). For the rest of this paper, we use the terms "reliability", "validity" and "quality" to refer to measurement reliability, validity and quality.

Two main approaches can be used to get this information: 1) estimate the quality of different survey questions by performing a Multitrait-Multimethod (MTMM) experiment, and 2) predict the quality of survey questions based on their characteristics using the Survey Quality Predictor (SQP) software (Saris, 2013). Both are explained in more details in section 2.

A lot of previous research has already been done to estimate the quality of different question's formats using MTMM experiments, starting with Andrews in 1984, and followed by many others (e.g. Költringer, 1995; Pan, 2015; Revilla, Saris, & Krosnick, 2014; Rodgers, Andrews, & Herzog, 1992; Saris, Revilla, Krosnick, & Shaeffer, 2010; Scherpenzeel, 2008; Scherpenzeel & Saris, 1997). However, most research has been done in face-to-face surveys or in the Dutch telepanel[2]. Nevertheless, nowadays, web surveys are more and more used, and the mode of data collection is one of the aspects that could influence the quality of survey questions. Indeed, the modes differ in terms of the presence or not of an interviewer and in the kind of stimuli (oral versus visual). Often, web surveys also ask questions in a more direct way (i.e. compared to the formal indirect way of asking in face-to-face surveys), and allow more diversity for the scales (e.g. drag and drop or sliders versus traditional rating scales). So, different levels of social desirability and measurement error are expected, as well as different levels of primacy, recency effects, etc. (De Leeuw, 2005).

Only few MTMM experiments have been conducted in web surveys. Scherpenzeel (2008) and Revilla and Saris (2013a) reported about some MTMM experiments included in the Dutch LISS probability-based panel. Revilla and Ochoa (2015), Revilla, Saris, Loewe and Ochoa (2015) and Revilla and Saris (2015) reported about different MTMM experiments implemented in the Netquest opt-in panels in Spain, Mexico and/or Colombia. Overall, these studies found usually high quality for the web survey data. When comparing it with other modes of data collection, they found that the quality is quite similar to the one of a face-to-face survey using visual aids, but significantly different of a telephone survey. Still, some differences are found between face-to-face and web quality, in some cases.

Furthermore, web surveys have evolved a lot in the last years, with the appearance of what has been called "unintended mobile respondents" (Peterson, 2012; Wells, Bailey, & Link, 2013), i.e. respondents who started to use mobile devices to answer to web surveys, even if this was not initially intended by the researchers. This mobile participation has been growing very quickly in different countries (Callegaro, 2010; De Bruijne & Wijnant, 2014; Revilla, Toninelli, Ochoa, & Loewe, 2016). The use of smaller and more mobile devices, could also affect the overall quality of web surveys. In the previously mentioned web survey studies

[2] Telepanel is a type panel where panellists are recruited to answer surveys through of computer assisted self-administered interviews (CASI). It was developed in the Netherlands in the 90's where the respondents were provided with an equipment allowing them to answer surveys from home without interviewer (Saris et al., 1998).

based on MTMM experiments, mobile completion was still almost inexistent. But now, just a few years later, it cannot be neglected.

In addition, past research using MTMM in web surveys has been done only in a few countries. Nevertheless, we know that the quality can vary across regions and languages (Oberski, Saris, & Hagenaars, 2007; Saris & Gallhofer, 2014). It can also vary depending on the question's topic, and only few topics have been tested in web surveys. Thus, more research estimating the quality of survey questions through MTMM experiments in the case of online surveys is needed, using more recent data, in different countries and for different topics.

Similarly, previous research using SQP to evaluate the quality of survey questions is limited. Some research has used it to evaluate face-to-face questions (e.g. Coromina & Saris, 2009; Coromina, Saris, & Oberski, 2008; Guillen, Coromina, & Saris, 2011; Revilla, Zavala-Rojas, et al., 2016; van der Zouwen & Smit, 2004). However, there is no research, to our knowledge, using the SQP approach to evaluate the quality of web survey questions.

To start filling in this gap, the aim of this study is to use these two different approaches to evaluate the quality of questions implemented in a web survey. Because both approaches have advantages and limitations, the aim is not only to evaluate quality, but also to compare them and assess whether SQP is an appropriate tool to evaluate web survey questions.

The rest of this paper is organized in the following way: Section 2 explains more in details the two methodologies used to assess quality. Section 3 presents the experimental questions for which we want to evaluate quality. Section 4 introduces the data and analyses conducted. Section 5 summarizes and compares the results obtained by the two approaches. Section 6 discusses the advantages and disadvantages of each approach to get information about survey instruments' quality. Finally, Section 7 concludes.

2. Evaluation of measurement quality

2.1. The Multitrait-Multimethod (MTMM) approach

Back in 1959, Campbell and Fiske proposed the MTMM design for the first time, suggesting that in order to study convergent and discriminant validity it is necessary to repeat a set of questions measuring correlated concepts of interest (called traits) using different methods (for instance response scales). Since then, this idea has been used as a basis to propose new ways to estimate the quality of survey questions. Different models have been developed for analysing MTMM data (Wothke, 1996), in particular confirmatory factor analysis models (Althauser, Heberlein, & Scott, 1971; Alwin, 1974; Andrews, 1984; Jöreskog, 1970, 1971; Werts & Linn, 1970).

In this paper, we use the True Score (TS) Model (Saris & Andrews, 1991) because a) this model provides better fit compared to others (Corten et al., 2002; Saris & Aalberts, 2003) and b) it allows estimating separately reliability, validity, method effect and the residual errors.

A TS-MTMM model with three traits and three methods is represented in Figure 1.

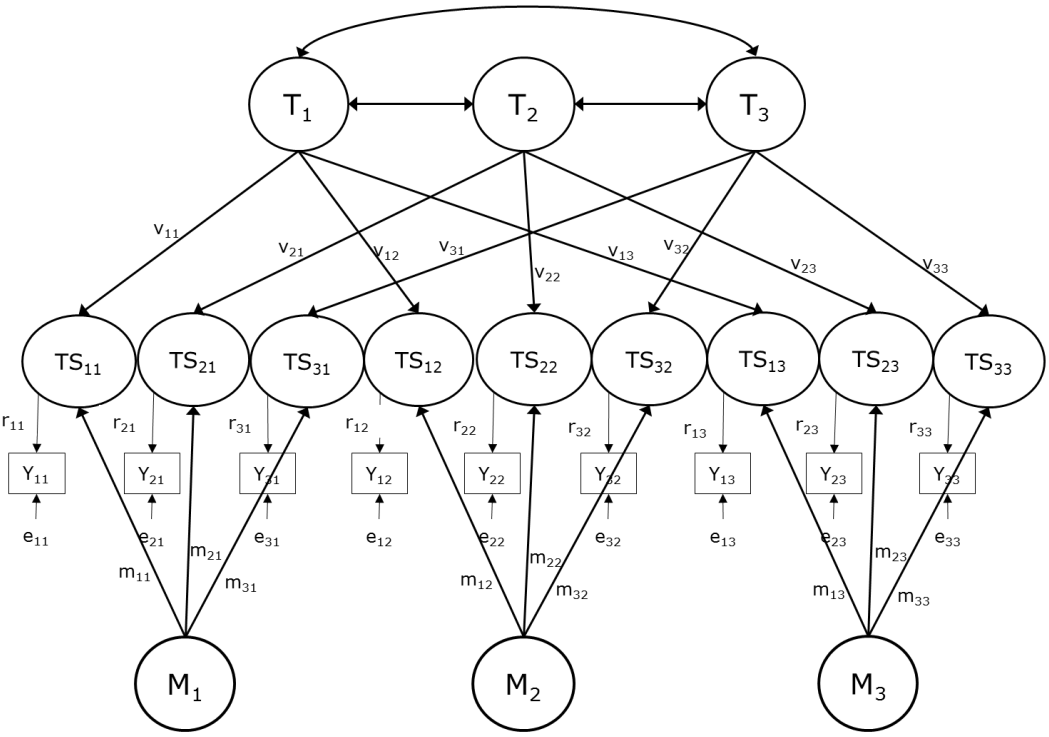

Figure 1: A True Score MTMM model with tree traits and three methods

T_i is the i^{th} latent variable of interest or trait; M_j is the j^{th} method factor. Y_{ij} is the observed variable for the i^{th} trait and the j^{th} method; TS_{ij} is the systematic component or true score of the response to Y_{ij}; e_{ij} is the random error associated with Y_{ij}. The effects $r_{ij,}$ v_{ij} and m_{ij} are respectively the reliability, validity and method effect coefficients for the i^{th} trait and the j^{th} method.

Following this model, reliability (r_{ij}^2) is defined as the strength of the relationship between the observed variable (Y_{ij}) and the true score (TS_{ij}), and is computed as the squared of the reliability coefficient. Validity (v_{ij}^2) is defined as the strength of the relationship between the true score (TS_{ij}) and the trait (T_i), and is computed as the squared of the validity coefficient. The quality (q_{ij}^2) represents the strength of the relationship between the observed variable (Y_{ij}) and the trait (T_i) and is computed as the product of reliability and validity. Reliability, validity and quality take values between 0 and 1. The closer to one, the better the measurement instrument is.

The model in Figure 1 assumes first, that the traits are correlated with each other; second, that the random errors are not correlated with each other, nor with the independent variables in the different equations; third, that the method factors are not correlated with each other, nor with the traits or the random errors; and forth, that the method effects for one specific method factor are equal for the different true scores. Estimates for all parameters of the model can be obtained using structural equation modelling software (e.g. LISREL, EQS or Mplus).

A limitation of this approach is that to identify a TS-MTMM model, usually three questions need to be repeated for the same respondents using at least three methods. This increases the cognitive burden of the respondents and threatens the accuracy of the measurements because

of memory effects, i.e. respondents might remember their previous answers and methods would be no longer independent. To avoid memory effects, van Meurs and Saris (1990) suggested, based on an experiment implemented in a face-to-face survey, that at least 20 minutes of similar questions should separate one question from its repetition, resulting in very long and costly questionnaires.

To lower the cognitive burden of the respondents linked to the repetitions, and reduce the questionnaires' length, Saris, Satorra and Coenders (2004) proposed combining the MTMM and the Split-Ballot approaches, resulting in the so-called Split-Ballot Multitrait-Multimethod (SB-MTMM) design. This consists in randomly assigning respondents to different groups. Each group gets the same questions asked with only two different methods but, each group gets a different combination of methods. Thus, the number of repetitions is reduced for each respondent and, at the same time, information about the three methods is obtained and the TS-MTMM model is identified under quite general conditions.

2.2. The Survey Quality Predictor (SQP) software

The MTMM approach has some major limitations. First, it requires supplementary data collection beyond the single survey question under evaluation, as well as additional data analysis. Thus, it is costlier, both in terms of money and time. Second, it is a post-hoc test, meaning that, when measurement error turns out to be high, it is already too late to improve the survey instrument in a given survey. The information can only be used for a later survey. Third, and most important, the MTMM approach can only be used for a subset of questions in each survey. It is clearly not possible, in practice, to repeat all questions of all surveys twice to get information about their quality. However, to what extent the information obtained for one survey instrument can be generalized to other survey instrument is still an open question.

Therefore, Andrews (1984) proposed to try to explain the information about quality from MTMM experiments by the questions' characteristics (p. 436). If quality can be well explained by questions' features, then, this information can be used to predict the quality of new questions.

This idea has been implemented by Saris, van der Veld and Gallhofer (2000) who launched the first SQP version in 2001. In 2012, it was further improved and a new version, SQP 2 (Saris et al., 2011), was made available for free to all users at: sqp.upf.edu.

The SQP software allows predicting the quality of survey questions by the identification of questions' features following a detailed coding scheme with up to 60 different formal and linguistic characteristics: for instance, the number of points in the response scale, the use of labels, the use of balanced or unbalanced questions, the polarity of questions, the presence of an introduction, the respondents' and interviewers' instructions, the mode of administration, or the position in the questionnaire.

SQP predicts the quality of survey questions based on a meta-analysis model that uses more than 3,000 MTMM estimates obtained from multiple surveys, for more than two decades, and in more than 20 different countries and languages (Oberski et al., 2011).

The major limitation of SQP is that the quality of the predictions depends on the data available for this huge meta-analysis of MTMM studies on which it is based. So far, most MTMMs were based in face-to-face and telepanel questions, not on web-surveys (although they are to some extent similar to telepanel studies). In this paper, to study if SQP can be used to evaluate the quality of web survey questions, we will compare the results from both approaches.

3. The experimental questions for which we want to evaluate measurements' quality

In this paper, we are interested in evaluating the quality of traditional questions about political satisfaction and trust in institutions. These concepts have been measured for many years in several large international surveys, like the European Social Survey, the Eurobarometer or the World and European Values Surveys. The questions asked to measure these concepts are largely used in substantive analyses to study and understand citizens' perceptions of national politics and institutions (e.g. Harrebye & Ejrnæs, 2015; Kaariainen, 2007; Linde & Ekman, 2003; Lühiste, 2014; O'Sullivan, Healy, & Breen, 2014; Shlapentokh, 2006; van der Meer, 2010; Zmerli & Newton, 2008).

Table 1 provides a summary of the content of the experimental questions studied in this paper. Three direct questions are about the *Norwegian political satisfaction* and three other indirect questions are about the level of *Trust in Norwegian institutions*. Each set of three questions belong to one experiment.

Table 1: Wording of questions for which we want to evaluate the quality

Topic	Trait	Wording of the questions
Norwegian political satisfaction	*Economy*	- How satisfied are you with the present state of the economy in Norway?
	Government	- How satisfied are you with the way the Norwegian government is doing its job?
	Democracy	- How satisfied are you with the way democracy works in Norway?
Trust in Norwegian institutions	*Parliament*	How high is your trust in the following institutions: - The parliament?
	Judiciary	- The judiciary?
	Police	- The police?

To identify the TS-MTMM model, we need to repeat each of these questions using three different scales. (from now on, we will call them "methods"). Table 2 shows the three methods used to measure the theoretically bipolar concept of satisfaction, and the three methods used to measure the theoretically unipolar concept of trust.

Table 2: Answer scales for which we want to evaluate the quality

	Norwegian political satisfaction
Method 1	o Very satisfied o Satisfied o Somewhat satisfied o Slightly satisfied

	○ Not satisfied at all										
Method 2	0 Extremely dissatisfied	1	2	3	4	5	6	7	8	9	10 Extremely satisfied
	○	○	○	○	○	○	○	○	○	○	○
Method 3	0 Very dissatisfied	1	2	3	4	5	6	7	8	9	10 Very satisfied
	○	○	○	○	○	○	○	○	○	○	○
	Trust in Norwegian institutions										
Method 1	Very high trust		High trust		Some trust		Low trust		No trust at all		
	○		○		○		○		○		
Method 2	0 No trust at all	1	2	3	4	5	6	7	8	9	10 Complete trust
	○	○	○	○	○	○	○	○	○	○	○
Method 3	No trust at all								Complete trust		
	○		○		○		○		○		

These methods differ at several levels: the polarity of the scale (bipolar or unipolar), the number of answer options (5-point scales or 11-point scales), the visual presentation (vertical or horizontal), the labelling (partially or fully labelled), the order of the options (from negative to positive or from positive to negative), the use of numbers next to textual labels or not, and the number of fixed reference points (like 'extremely' or 'completely' instead of 'very'). In the case of the trust experiment, the questions are always presented in a grid format (or battery).

4. Data and analysis

4.1. For the MTMM experiments

4.1.1. Data from the Norwegian Citizen Panel (NCP)

The NCP started in 2012 and is a research-purpose internet panel with more than 10,000 panel members. A probability-based sample of the general Norwegian population from 18 to 95 years old was drawn from the Norwegian National Registry. Panel members complete an online questionnaire of about 20 minutes two times a year[3].

All the questions presented in Table 1 measured with the different methods presented in Table 2, form the two MTMM experiments (of three traits each) asked in the 5th NCP wave. The data collection took place from October to November 2015. After the invitation email and three reminders, the overall response rate was 62%. Over 5,000 panellists participated in this wave. The NCP 5th wave data[4] included over 500 variables. Participants were divided in four subsets that received different questions. On average, the respondents took between 20 to 25

[3] For more information about the NCP, we refer to: http://digsscore.uib.no/methodology.

[4] "Norwegian Citizen Panel Wave 5, 2015". Data collected by Ideas 2 Evidence for Elisabeth Ivarsflaten, University of Bergen. First NSD edition, Bergen 2015.

minutes to complete the questionnaire. Different experiments were included, among which these two MTMM experiments to evaluate the reliability and validity of the questions presented above. The non-response rate in those questions was, on average, 1.7%. Data on the type of device (i.e. smartphone, tablet or PC) used to answer the survey was not collected and questions were not adapted to the different devices (Skjervheim & Høgestøl, 2015). This set of experimental questions was answered by a subset of 1,277 panellists following the split-ballot design presented in Table 3.

Table 3: 3-group split-ballot MTMM design used in this study (N=1,277)

	Time 1	Time 2	N
Group 1	Method 1	Method 2	413
Group 2	Method 1	Method 3	442
Group 3	Method 2	Method 3	422

Group 1 received method 1 at time 1 and method 2 at time 2. Group 2 received method 1 at time 1 and method 3 at time 2. Finally, group 3 received method 2 at time 1 and method 3 at time 2. There were about 50 questions between time 1 and time 2. However, response times were not registered so we cannot compute how much time separated one experimental question from its repetition. In a web survey, having information about the individual response time is crucial to evaluate the time between responses because it can vary a lot across respondents. Some respondents speed through the survey (Zhang & Conrad, 2013), or are quicker in reading than others. Some respondents can also stop and come back later. Thus, a very short time as well as a very long time can separate repetitions between these experimental questions too. This is not desirable either because opinions may change.

Overall, the questionnaire counted about 80 questions about national politics, climate change, attitudes towards refugees and a monetary experiment. An English translation of the MTMM questions studied in this paper is provided in Appendix A.

4.1.2. MTMM analyses and testing

We analysed each of the two SB-MTMM experiments implemented in the NCP separately. For each experiment, the estimates were obtained using the LISREL 8.72 software with Maximum Likelihood estimation for multi-group analysis based on the covariance matrix, means and standard errors (Jöreskog & Sörbom, 1996). The estimation of both initial models, (cf. Appendix B) resulted in improper solutions, i.e. negative variances. This is quite common in such type of models, as shown by Revilla and Saris (2013b). To get a proper solution, we allowed a correlation between methods 2 and 3 in the political satisfaction experiment. In the trust in institutions experiment, we allowed the effect of method 2 on trait 2 to be different from the other method effects.

Once a proper solution was found, the model was tested for misspecifications. Hu and Bentler (1998) stated that ‘a model is said to be misspecified when (a) one or more parameters are estimated whose population values are zeros (i.e., an over-parameterized misspecified model), (b) one or more parameters are fixed to zeros whose population values are non-zeros (i.e., an under-parameterized misspecified model) or both’ (p. 427). In addition, a model can also be misspecified when (c) one or more parameters are fixed to a certain value which is different

from its population value, or (d) when one or more parameters are constraint to be equal across groups when their true values are in fact different.

For the global fit of the model, we considered the chi-square test and Root Mean Square Error of Approximation (RMSEA) test. However, since these global tests of model fit have problems (Saris, Satorra, & van der Veld, 2009), we mainly focused on the local fit to detect if there were misspecified parameters. To do so, we used the JRule software version 3.0.4 (van der Veld, Saris, & Satorra, 2008), which is based on the procedure for testing model misspecifications of each restricted parameter using expected parameter changes, modification indices and the power, as proposed by Saris, Satorra and van der Veld (2009).

JRule provides a test for misspecifications at the parameter level. The minimum size of the misspecification that we wanted to detect in the restricted parameters was fixed at 0.1. Necessary corrections were introduced in the two models until an acceptable global fit was obtained. The summary of the final model adjustments is provided in Appendix C.

Since respondents were randomly assigned to the split-ballot groups, we usually do not expect differences between variables with the same method asked to different groups. However, we can expect to find differences for questions using method 2 because in group 1 they were asked at time 2 and in group 3 at time 1. On the one hand, if respondents get tired at the end of the survey, this can lead to lower quality at time 2. On the other hand, if respondents remember their previous answer at time 2, this can lead to higher quality. That is why, in Table 4, different estimates are provided for method 2 for the different points in time. The results of this analysis are presented in section 5.1.

4.2. For the SQP predictions

The quality predictions of the 18 questions, the nine questions per experiment presented in section 3, were obtained using the latest version of SQP, 2.1. Moreover, the questions for method 2 were coded twice to consider the position in the questionnaire: in group 1 method 2 was asked at time 1 while in group 3 it was asked at time 2. Therefore, a total of 24 questions were coded.

For each of the 24 questions, the characteristics of the questions in Norwegian were coded following the SQP 2.1 Coding Instructions[5]. These characteristics are related with the topic, the formulation of the request, the type of response scale used, the use of instructions for the respondent or the interviewer, the complexity of the question, the visual presentation and the mode of data collection were coded. SQP allows to identify questions asked in web surveys by indicating that the questions were administrated using a computer and without the presence of an interviewer.

After all the characteristics of the questions are coded, SQP provides information about the quality obtained based on its prediction algorithm. Thus, no extra analyses are needed with this approach.

[5] Downloaded from: http://sqp.upf.edu/media/files/sqp_coding_instructions.pdf

The questions, the codes and the resulting quality predictions are stored in the SQP 2.1 database[6]. The predictions obtained from this process are shown in section 5.2.

5. Results

This section provides the quality estimates and predictions obtained using both approaches. With this information, we can evaluate the questions used in the 5th wave of the NCP, and compare both approaches to assess the feasibility of the SQP prediction for online surveys.

5.1. MTMM estimates of measurement quality

First, the quality obtained through the MTMM analyses is presented in Table 4 for each experiment, trait and method.

Table 4: MTMM estimates of quality

Experiment	Traits	Method 1 Time 1	Method 2 Time 1	Method 2 Time 2	Method 3 Time 2
Norwegian political satisfaction	*Economy*	.60	.76	.81	.61
	Government	.85	.81	.85	.73
	Democracy	.74	.85	.89	.63
Trust in Norwegian institutions	*Parliament*	.68	.81		.73
	Judiciary	.72	.80		.73
	Police	.72	.88		.76

Overall, both experiments provide similar levels of quality; in the *Norwegian political satisfaction* experiment the average quality is 0.74, while in the *Trust in Norwegian institutions* it is 0.76. However, the MTMM quality estimates of the questions tested varies across traits and methods, from a minimum of 0.60 (Economy, Method 1, Time 1) to a maximum of 0.89 (Democracy, Method 2, Time 2).

In both experiments, the highest quality is observed for method 2, a horizontal scale with 11 points, partially labelled, ordered from negative to positive and with two fixed reference points.

Comparing methods 1 and 3, in the political satisfaction experiment, we observe a higher quality for method 1, a vertical 5-point, fully labelled scale, ordered from positive to negative with one fixed reference point, than for method 3, a horizontal 11-point, partially labelled scale, ordered from negative to positive, with no labelled fixed reference point. When looking at the results of the trust in institutions experiment, however, we do not see significant differences between method 1 and method 3 (this time, a 5-point battery scale, partially labelled, ordered from negative to positive and with two fixed reference points), although, in this case, quality is slightly higher in method 3.

[6] All the information is available in SQP in a study named "Norwegian Citizen Panel W5" which can be consulted by any SQP user. Everybody can become an SQP user just by registering.

In Table 4, we can see that method 2 obtained higher quality at time 2 than at time 1. That suggests that memory effect might be present. If respondents, at time 2, remember their previous answer, this indeed can lead to higher quality estimates.

Overall, based on these results, we would recommend for future waves of the NCP to use 11-point partially labelled scales, ordered from negative to positive and with fixed reference points, to measure questions about satisfaction and trust.

5.2. SQP prediction of measurement quality

Second, the quality predicted using SQP, by coding the characteristics of the questions in Norwegian is presented in Table 5.

Table 5: SQP 2.1 quality predictions

Experiment	Traits	Method 1 Time 1	Method 2 Time 1	Method 2 Time 2	Method 3 Time 2
Norwegian political satisfaction	*Economy*	.70	.74	.67	.67
	Government	.70	.74	.67	.66
	Democracy	.70	.74	.66	.66
Trust in Norwegian institutions	*Parliament*	.69	.72	.67	.61
	Judiciary	.73	.78	.75	.71
	Police	.73	.79	.75	.71

Overall, both experiments provided similar levels of quality; in the *Norwegian political satisfaction* experiment the average quality is 0.69 while in the *Trust in Norwegian institutions* it is 0.72. However, the quality predictions vary across traits and methods, from 0.61 (Parliament, Method 3, Time 2) to 0.79 (Police, Method 2, Time 1).

Again, the higher quality is observed in both experiments for method 2. Moreover, in both experiments, method 1, performs better than method 3.

In this case, the predictions obtained in method 2 are lower at time 2, compared to time 1, which is different from what we found using the MTMM estimates. The lower quality at time 2 suggests now an increased fatigue, leading to higher random errors, which lower the quality of items placed towards the end of the survey.

5.3. Comparison between MTMM estimation and SQP prediction

While the MTMM approach gives us an estimation of the quality for a set of specific questions under evaluation in a specific survey and sample, the SQP approach predicts the quality based on the cumulative knowledge obtained by combining, in a meta-analysis, the information about the quality of thousands of MTMM questions and its characteristics. These two approaches have advantages and disadvantages as discussed in section 2.

Comparing the results of Tables 4 and 5, we see first that the SQP predictions and the MTMM estimates for method 1, method 2 at time 1 and method 3, have an average difference lower

than 0.1. Second, we see that the MTMM qualities for method 2 at time 2 are overestimated for approximately 0.15 compared to the SQP predictions. This difference could be explained by several factors. On the one hand, the MTMM estimates may be biased since a too short time in between repetitions can lead respondents to remember their previous answer at time 2. On the other hand, SQP aims to predict the quality of questions which are asked for the first time in a questionnaire, since questions are normally only asked once. Therefore, it does allow us predicting the quality of a repetition where some memory effects might have occurred. Instead, the SQP predictions consider the position of the items in the questionnaire. Usually items more at the end of the questionnaire have a lower quality because respondents may get tired. This difference between time 1 and 2 for method 2 is only interesting in methodological terms. For future uses of the qualities reported here, we suggest using time 1 for method 2. From now on, we will focus only on the estimates for time 1.

Overall, the MTMM estimates are on average 0.043 higher than the SQP predictions (with 0.010 of minimum difference and 0.15 of maximum). Table 6 presents the correlation between the SQP predictions and the MTMM estimates, for reliability, validity and quality.

Table 6: Correlation between SQP predictions and MTMM estimates for reliability (r^2), validity (v^2) and quality (q^2)

	Reliability	Validity	Quality
Pearson Correlation between MTMM and SQP	.74***	.82***	.61***

Significance threshold: p<0.01 ***

Correlations between MTMM estimates and SQP predictions, for the three indicators are quite high and significant. The highest correlation is between the validity estimates from the MTMM analyses and the SQP predictions, followed by the reliability and then, the quality.

6. Discussion about the two approaches

Having presented and compared the results from the two alternatives, two questions remain.

First, can SQP be used to evaluate web survey questions? Our findings suggest that it can be used for scales that do not differ too much from those included in the SQP meta-analysis[7]. Thus, it can be useful for typical scales used in face-to-face surveys or in telepanels, like the ones tested in this study. However, there are some new scales (e.g. drag and drop scales) that have been developed for web surveys which cannot be coded properly in SQP yet. In this case, MTMM experiments need to be implemented to estimate the quality of new forms of questions.

Second, when both an MTMM estimate and an SQP prediction can be obtained, what approach should be preferred? This depends. While the MTMM are based on actual data from the responses obtained to different experimental questions, the SQP predictions are based on a meta-analysis that relates the cumulative knowledge about the quality from a large amount of MTMMs to the characteristics of survey questions. The SQP predictions, therefore, are an

[7] For more information about the type of questions included in the current version of SQP see https://drive.google.com/open?id=0B9vo3n40fqoFdG1NOFNjY2hnU3c.

evaluation tool available a priori when designing a questionnaire. As such, they can aid already at the survey design stage in taking decisions of what type of scales or question formulations should be preferred to achieve high data quality (Revilla, Zavala-Rojas, et al., 2016). On the contrary, MTMM estimates are only available after data collection. But to correct relationships from measurement error in a study, we would in principle suggest using the quality estimates obtained from the MTMM experiments in that particular study, since they will be specific of that study. However, when choosing the MTMM approach one should take into account that non-convergence and negative variances in these kinds of MTMM models are frequent and, in practice, often difficult to solve (Revilla and Saris, 2013b).

Thus, even when MTMM data is available for a study, we recommend always using the SQP predictions to validate the results found with the MTMM approach. On the one hand, if both provide very similar results, one can then be quite confident in using the MTMM estimates to correct for measurement error. On the other hand, if MTMM estimates and SQP predictions differ significantly, we suggest comparing the substantive results of interest (e.g. regression coefficients) after correction for measurement error obtained using the two different approaches. Researchers should then decide upon what results make more sense, keeping in mind the uncertainty.

7. Conclusions, limitations and further research

The goal of this paper was to use two different approaches to evaluate the measurement quality of a set of questions implemented in a web survey and assess whether SQP is an appropriate tool to evaluate web survey questions. Thus, first we conducted two MTMM experiments in the 5th wave of the NCP. Each experiment included three different traits measured by three different methods. The administration of the different methods was randomized using a split-ballot design; so, each respondent only had to answer to the same question twice. Second, we used the SQP prediction algorithm to obtain a prediction of the quality of these same questions based on their characteristics. Finally, we compared the results obtained and discussed the appropriateness of each approach.

We found that on average the quality of the questions included in the MTMM experiments is 0.75. This is quite high compared to what is often observed in practice. However, on average, still 25% of the observed variance is due to measurement error, indicating that correction for measurement errors would be necessary. For all six traits studied, the MTMM experiments provided evidence in favour of the use of the partially labelled 11-point scale with fixed reference points and ordered from negative to positive.

These MTMM results are, however, not exempt of limitations. First, because of the short extent of the survey (i.e. between 20 to 25 minutes on average), the results between questions at time 1 and time 2 may be biased because of the presence of memory effects. Timing variables were not available in this study. Thus, we could not control if enough time separated the repetitions. Second, the results are specific to the set of experimental questions evaluated in the 5th wave of the Norwegian Citizen Panel, for the Norwegian population in 2015 and they should not be generalized to other countries, languages or to other types of questions.

Thus, in this paper, and for the first time, the MTMM estimates are compared to the SQP predictions for a web survey. Even if SQP is currently not based on data from web surveys

and besides some differences, we found that in general the SQP predictions obtained are in line to what we found using the MTMM estimation approach. This suggests that SQP predictions can be used for web surveys at least for some topics and scales (e.g. radio button scales as investigated here). The general conclusions do not change. Therefore, overall, we can be quite confident about the results obtained with both approaches.

Moreover, SQP allows to evaluate the quality of questions and/or to correct for measurement error without the need to collect extra data. However, collecting MTMM data can still be useful to provide a more specific evaluation of the quality of the questions under specific circumstances. To control for possible biases coming from the MTMM analytical problems, we recommend making corrections for measurement errors using the quality information provided by both approaches (i.e. as a kind of sensitivity analysis).

The SQP approach is also not exempt of limitations. SQP allows predicting the quality of survey questions in a large range of languages (21), but still many other languages are not available. It also allows predicting the quality for many different scales formats, but still some new forms of scales (e.g. drag and drop or order by click) are not available.

To conclude, several points still need to be further investigated. First, the necessary timing in between repetitions of the experimental questions, to avoid memory effects in web surveys. Second, more evidence towards the quality of online survey instruments using different topics and different methods is needed, especially for those newest types of scales commonly used in web surveys. Third, further research should also consider differences between the devices used to respond to the online survey, i.e. smartphone, tablet or PC. Finally, to make SQP more useful for any kind of research, further research should be directed to improve the SQP quality predictions and extend its possibilities to new modes of asking questions.

References

Althauser, R. P., Heberlein, T. A., & Scott, R. A. (1971). A causal assessment of validity: The augmented mulitraitmultimethod matrix. In H. M. Blalock Jr. (Ed.), *Causal Models in the Social Sciences* (pp. 151–169). Chicago: Aldine.

Alwin, D. F. (1974). An analytic comparison of four approaches to the interpretation of relationships in the multitrait-multimethod matrix. In H. L. Costner (Ed.), *Sociological Methodology* (pp. 79–105). San Francisco: Jossey Bass.

Andrews, F. M. (1984). Construct validity and error components of survey measures: a structural modelling approach. *Public Opinion Quarterly*, *48*(2), 409–442. http://doi.org/10.1086/268840

Borgatta, E. F., & Bohrnstedt, G. W. (1980). Level of Measurement: Once Over Again. *Sociological Methods & Research*, *9*(2), 147–160. http://doi.org/10.1177/004912418000900202

Callegaro, M. (2010). Do You Know Which Device Your Respondent Has Used to Take Your Online Survey? *Survey Practice*.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrices. *Psychological Bulletin*, *56*(2), 81–105.

Coromina, L., & Saris, W. E. (2009). Quality of media use measurement. *International Journal of Public Opinion Research*, *21*(4), 424–450. http://doi.org/10.1093/ijpor/edp014

Coromina, L., Saris, W. E., & Oberski, D. L. (2008). The Quality of the Measurement of Interest in the Political Issues presented in the Media in the ESS. *ASK. Research and Methods*, *17*, 7–38.

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

De Bruijne, M., & Wijnant, A. (2014). Improving Response Rates and Questionnaire Design for Mobile Web Surveys. *Public Opinion Quarterly*, *78*(4), 951–962. http://doi.org/10.1093/poq/nfu046

De Leeuw, E. D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, *21*(2), 233–255.

DeCastellarnau, A., & Saris, W. E. (2014). A simple way to correct for measurement errors in survey research. Retrieved from http://essedunet.nsd.uib.no/cms/topics/measurement/

Garner, W. R. (1960). Rating scales, discriminability, and information transmission. *Psychological Review*, *67*, 343–52.

Guillen, L., Coromina, L., & Saris, W. E. (2011). Measurement of Social Participations and its Place in Social Capital Theory. *Social Indicators Research*, *100*(2), 331–350. http://doi.org/10.1007/s11205

Harrebye, S., & Ejrnæs, A. (2015). European patterns of participation – How dissatisfaction motivates extra-parliamentary activities given the right institutional conditions.

Comparative European Politics, *13*(2), 151–174. http://doi.org/10.1057/cep.2013.7

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model specificiation. *Psychological Methods*, *3*, 424–453.

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology1*, *23*(2), 121–145. http://doi.org/10.1111/j.2044-8317.1970.tb00439.x

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133. http://doi.org/10.1007/BF02291393

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's Reference Guide*. Uppsala, Sweden: Scientific Software International.

Kaariainen, J. T. (2007). Trust in the Police in 16 European Countries: A Multilevel Analysis. *European Journal of Criminology*, *4*(4), 409–435. http://doi.org/10.1177/1477370807080720

Költringer, R. (1995). Measurement quality in Austrian personal interview surveys. In W. E. Saris & Á. Münnich (Eds.), *The Multitrait-Multimethod Approach to evaluate measurement instrument* (pp. 207–225). Budapest: Eötvös University Press.

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing Rating Scales for Effective Measurement in Surveys. In L. E. Lyberg, P. P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey Measurement and Process Quality* (pp. 141–164). Hoboken, NJ.

Linde, J., & Ekman, J. (2003). Satisfaction with democracy: A note on a frequently used indicator in comparative politics. *European Journal of Political Research*, *42*(3), 391–408. http://doi.org/10.1111/1475-6765.00089

Lühiste, K. (2014). Social Protection and Satisfaction with Democracy: a Multi-level Analysis. *Political Studies*, *62*(4), 784–803. http://doi.org/10.1111/1467-9248.12080

O'Sullivan, S., Healy, A. E., & Breen, M. J. (2014). Political Legitimacy in Ireland During Economic Crisis: Insights from the European Social Survey. *Irish Political Studies*, *29*(4), 547–572. http://doi.org/10.1080/07907184.2014.942645

Oberski, D. L., Gruner, T., & Saris, W. E. (2011). The prediction procedure of the quality of the questions based on the present data base of questions. In W. E. Saris, D. L. Oberski, M. Revilla, D. Zavala-Rojas, L. Lilleoja, I. N. Gallhofer, & T. Gruner (Eds.), *RECSM Working-Paper N°24* (pp. 71–88). Barcelona.

Oberski, D. L., Saris, W. E., & Hagenaars, J. A. P. (2007). Why are there differences in measurement quality across countries? In G. Loosveldt & Swyngedouw (Eds.), *Measuring Meaningful Data in Social Research* (pp. 281–300). Leuven: Acco.

Pan, M. (2015). Rating scale validation: An MTMM approach. In *Nonverbal delivery in speaking assessment. From an argument to a rating scale formulation and validation* (pp. 119–214). Singapore: Springer. http://doi.org/10.1007/978-981-10-0170-3_7

Peterson, G. (2012). Unintended mobile respondents. In *Paper presented at the Annual Council of American Survey Research Organizations Technology Conference* (pp. 1–31). New York, NY.

Revilla, M., & Ochoa, C. (2015). Quality of Different Scales in an Online Survey in Mexico and Colombia. *Journal of Politics in Latin America*, *7*(3), 157–177.

Revilla, M., & Saris, W. E. (2013a). A Comparison of the Quality of Questions in a Face-to-face and a Web Survey. *International Journal of Public Opinion Research*, *25*(2), 242–253. http://doi.org/10.1093/ijpor/eds007

Revilla, M., & Saris, W. E. (2013b). The Split-Ballot Multitrait-Multimethod Approach: Implementation and Problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 27–46.

Revilla, M., & Saris, W. E. (2015). Estimating and comparing the quality of different scales of an online survey using an MTMM approach. In U. Engel (Ed.), *Survey Measurements: Techniques, Data Quality and sources of Error* (pp. 53–74). Frankfurt, New York: Campus Verlag.

Revilla, M., Saris, W. E., & Krosnick, J. A. (2014). Choosing the Number of Categories in Agree-Disagree Scales. *Sociological Methods & Research*, *43*(1), 73–97. http://doi.org/10.1177/0049124113509605

Revilla, M., Saris, W. E., Loewe, G., & Ochoa, C. (2015). Can a non-probabilistic online panel achieve question quality similar to that of the European Social Survey? *International Journal of Market Research*, *57*(3), 395–412.

Revilla, M., Toninelli, D., Ochoa, C., & Loewe, G. (2016). Do online access panels need to adapt surveys for mobile devices? *Internet Research*, *26*(5), 1209–1227. http://doi.org/10.1108/IntR-02-2015-0032

Revilla, M., Zavala-Rojas, D., & Saris, W. E. (2016). Creating a good question: How to use cumulative experience. In C. Wolf, D. Joye, T. W. Smith, & Yang-Chih Fu (Eds.), *The SAGE-Handbook of Survey Methodology* (pp. 236–254). SAGE.

Rodgers, W. L., Andrews, F. M., & Herzog, A. R. (1992). Quality of survey measures: a structural modeling approach. *Journal of Official Statistics*, *8*(3), 251–275.

Saris, W. E. (2013). The prediction of question quality: the SQP 2.0 software. In B. Kleiner, I. Renschler, B. Wernli, P. Farago, & D. Joye (Eds.), *Understanding Research Infrastructures in the Social Sciences* (pp. 135–144). Zurich: Seismo Press.

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

Saris, W. E., & Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 575–598). New York: John Wiley and Sons, Inc.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ: John Wiley and Sons, Inc. http://doi.org/0.1002/9780470165195

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

Saris, W. E., Oberski, D. L., Revilla, M., Zavala-Rojas, D., Lilleoja, L., Gallhofer, I. N., & Gruner, T. (2011). *The development of the program SQP 2.0 for the prediction of the quality of survey questions* (RECSM Working Paper No. 24). Barcelona.

Saris, W. E., & Revilla, M. (2016). Correction for measurement errors in survey research: necessary and possible. *Social Indicators Research*, *127*(3), 1005–1020. http://doi.org/10.1007/s11205-015-1002-x

Saris, W. E., Revilla, M., Krosnick, J. A., & Shaeffer, E. M. (2010). Comparing Questions with Agree/Disagree Response Options to Questions with Item-Specific Response Options. *Survey Research Methods*, *4*(1), 61–79.

Saris, W. E., Satorra, A., & Coenders, G. (2004). A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design. *Sociological Methodology*, *34*(1), 311–347. http://doi.org/10.1111/j.0081-1750.2004.00155.x

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*. http://doi.org/10.1080/10705510903203433

Saris, W. E., van der Veld, W. M., & Gallhofer, I. N. (2000). *A program for prediction of the Quality of Survey Measurement*. Amsterdam.

Saris, W. E., van Wijk, T., & Scherpenzeel, A. C. (1998). Validity and Reliability of Subjective Social Indicators. *Social Indicators Research*, *45*, 173–199.

Scherpenzeel, A. C. (2008). *Online interviews and data quality: A multitrait-multimethod study* (Draft paper to be presented at the MESS Workshop, 22-23 August 2008, Zeist.). Tilburg University.

Scherpenzeel, A. C., & Saris, W. E. (1997). The Validity and Reliability of Survey Questions: A Meta-Analysis of MTMM Studies. *Sociological Methods and Research2*, *25*(3), 341–383.

Shlapentokh, V. (2006). Trust in public institutions in Russia: The lowest in the world. *Communist and Post-Communist Studies*, *39*(2), 153–174. http://doi.org/10.1016/j.postcomstud.2006.03.004

Skjervheim, Ø., & Høgestøl, A. (2015). Norwegian Citizen Panel 2015, Fith wave: Methodology report. Retrieved from http://www.nsd.uib.no/data/individ/publikasjoner/NSD2343/NSD2343rapport.pdf

van der Meer, T. (2010). In what we trust? A multi-level study into trust in parliament as an evaluation of state characteristics. *International Review of Administrative Sciences*, *76*(3), 517–536. http://doi.org/10.1177/0020852310372450

van der Veld, W. M., Saris, W. E., & Satorra, A. (2008). Judgement rule aid for structural equation models version 3.0.4 beta.

van der Zouwen, J., & Smit, J. H. (2004). Evaluating Survey Questions by Analyzing Patterns of Behavior Codes and Question–Answer Sequences: A Diagnostic Approach, in Methods for Testing and Evaluating Survey Questionnaires. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 109–130). Hoboken, NJ, USA: John Wiley & Sons, Inc. http://doi.org/10.1002/0471654728.ch6

van Meurs, A., & Saris, W. E. (1990). Memory effects in MTMM studies. In W. E. Saris & A. van Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies* (pp. 134–146). Amsterdam: North Holland.

Wells, T., Bailey, J. T., & Link, M. W. (2013). Filling the Void: Gaining a Better Understanding of Tablet-based Surveys. *Survey Practice*, *6*(1), 1–4.

Werts, C. E., & Linn, R. L. (1970). Path analysis: Psychological examples. *Psychological Bulletin*, *74*(3), 193–212. http://doi.org/10.1037/h0029778

Wothke, W. (1996). Models for Multitrait-Multimethod Analysis. In G. C. Marcoulides & R. E. Schumacker (Eds.), *Advanced Structural Equation Modeling. Issues and Techniques* (pp. 7–56). Mahwah, NJ: Lawrence Erlbaum.

Zhang, C., & Conrad, F. (2013). Speeding in Web Surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*(2), 127–135. http://doi.org/10.18148/srm/2014.v8i2.5453

Zmerli, S., & Newton, K. (2008). Social Trust and Attitudes Toward Democracy. *Public Opinion Quarterly*, *72*(4), 706–724. http://doi.org/10.1093/poq/nfn054

Appendix A: The NCP Wave 5 SB-MTMM experiments' formulation

- **EXPERIMENT 1 – Norwegian political satisfaction**

Method 1/Trait 1: On the whole how satisfied are you with the present state of the economy in Norway?

- o Very satisfied
- o Satisfied
- o Somewhat satisfied
- o Slightly satisfied
- o Not satisfied at all

Method 1/Trait 2: Now thinking about the Norwegian government, how satisfied are you with the way it is doing its job?

- o Very satisfied
- o Satisfied
- o Somewhat satisfied
- o Slightly satisfied
- o Not satisfied at all

Method 1/Trait 3: And on the whole, how satisfied are you with the way democracy works in Norway?

- o Very satisfied
- o Satisfied
- o Somewhat satisfied
- o Slightly satisfied
- o Not satisfied at all

Method 2/Trait 1: On the whole how satisfied are you with the present state of the economy in Norway?

0 Extremely dissatisfied	1	2	3	4	5	6	7	8	9	10 Extremely satisfied
○	○	○	○	○	○	○	○	○	○	○

Method 2/Trait 2: Now thinking about the Norwegian government, how satisfied are you with the way it is doing its job?

0 Extremely dissatisfied	1	2	3	4	5	6	7	8	9	10 Extremely satisfied
○	○	○	○	○	○	○	○	○	○	○

Method 2/Trait 3: And on the whole, how satisfied are you with the way democracy works in Norway?

0 Extremely dissatisfied	1	2	3	4	5	6	7	8	9	10 Extremely satisfied
○	○	○	○	○	○	○	○	○	○	○

Method 3/Trait 1: On the whole how satisfied are you with the present state of the economy in Norway?

0 Very dissatisfied	1	2	3	4	5	6	7	8	9	10 Very satisfied
○	○	○	○	○	○	○	○	○	○	○

Method 3/Trait 2: Now thinking about the Norwegian government, how satisfied are you with the way it is doing its job?

0 Very dissatisfied	1	2	3	4	5	6	7	8	9	10 Very satisfied
○	○	○	○	○	○	○	○	○	○	○

Method 3/Trait 3: And on the whole, how satisfied are you with the way democracy works in Norway?

0 Very dissatisfied	1	2	3	4	5	6	7	8	9	10 Very satisfied
○	○	○	○	○	○	○	○	○	○	○

- **EXPERIMENT 2 – Trust in Norwegian institutions**

Method 1/Traits 1-3: How high is your trust in the following institutions?

	Very high trust	High trust	Some trust	Low trust	No trust at all
The parliament	○	○	○	○	○
The judiciary	○	○	○	○	○
The police	○	○	○	○	○

Method 2/Traits 1-3: Please indicate on a score of 0-10 how much you personally trust each of these institutions. 0 means you do not trust the institution at all and 10 means you have complete trust.

	0 No trust at all	1	2	3	4	5	6	7	8	9	10 Complete trust
The parliament	○	○	○	○	○	○	○	○	○	○	○
The judiciary	○	○	○	○	○	○	○	○	○	○	○
The police	○	○	○	○	○	○	○	○	○	○	○

Method 3 /Traits 1-3: Please indicate how much you personally trust each of these institutions.

	No trust at all				Complete trust
The parliament	○	○	○	○	○
The judiciary	○	○	○	○	○
The police	○	○	○	○	○

Appendix B: Example of a LISREL's initial model input

```
Analysis of NCP Wave 5 Satisfaction group 1
Data ng=3 ni=9 no=413 ma=cm
km
1
0.230106522195765 1
0.404913915084517 0.233841377035696 1
-0.683470833184798 -0.221926616056117 -0.391651969362765 1
-0.266572123503679 -0.843716030350421 -0.21250117153763 0.324181312389196 1
-0.480442388671978 -0.18024994439464 -0.800111264750463 0.561626808012355 0.227773981107847 1
0 0 0 0 0 0 1
0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 1

me
2.40831295843521 3.07960199004975 2.22413793103448 7.42857142857143 6.13432835820895 7.80246913580247 0.0
0.0 0.0

sd
0.850045754162623 0.917348363896813 0.867241188882013 1.79534966602064 2.17944733703399 1.87360346068256
1.0 1.0 1.0

model ny=9 ne=9 nk=6 ly=fu,fi te=di,fi ps=sy,fi be=fu,fi ga=fu,fi ph=sy,fi
value 1 ly 1 1 ly 2 2 ly 3 3 ly 4 4 ly 5 5 ly 6 6  !loadings fixed to 1 for identification purposes
value 0 ly 7 7 ly 8 8 ly 9 9
                                                    !some ly are fixed to 0 because they not apply to this split-ballot group
free te 1 1 te 2 2 te 3 3 te 4 4 te 5 5 te 6 6
value 1 te 7 7 te 8 8 te 9 9                    !some te are fixed to 0 because they not apply to this split-ballot group

free ga 1 1 ga 4 1 ga 7 1 ga 2 2 ga 5 2 ga 8 2 ga 3 3 ga 6 3 ga 9 3
value 1 ga 1 4 ga 2 4 ga 3 4 ga 4 5 ga 5 5 ga 6 5 ga 7 6 ga 8 6 ga 9 6  !these ga are the method effects, set to 1 following the
                                                                        !4th assumption.
free ph 2 1 ph 3 1 ph 3 2
free ph 4 4 ph 5 5 ph 6 6
value 1 ph 1 1 ph 2 2 ph 3 3  !variances fixed to 1 for identification purposes.

start 0.5 all

out mi iter=2000 adm=off sc ec

Analysis of NCP Wave 5 Satisfaction group 2
Data ni=9 no=442 ma=cm
Km
1
0.140094287671227 1
0.350159534430647 0.158551546221456 1
0 0 0 1
0 0 0 0 1
0 0 0 0 0 1
-0.594889524722528 -0.126549146761991 -0.331818310793426 0 0 0 1
-0.138984447809147 -0.792260608003386 -0.172671492671237 0 0 0 0.332863568754183 1
-0.351299559777779 -0.10049897179435 -0.721600725477894 0 0 0 0.510553380750539 0.267635034336383 1

me
2.4624145785877 3.06004618937644 2.23502304147465 0.0 0.0 0.0 7.51029748283753 6.09862385321101
7.97921478060046

sd
0.877327095385977 0.94089053447338 0.851608517126538 1.0 1.0 1.0 2.0121213310212 2.3355125329404
2.07432102513969

model ny=9 ne=9 nk=6 ly=fu,fi te=di,fi ps=in be=in ga=in ph=in
```

!assuming that split-ballot groups are randomized, we do not expect differences between groups, so they are set to be invariant, except for those parameters affected by the randomization of the methods.
value 1 ly 1 1 ly 2 2 ly 3 3 ly 7 7 ly 8 8 ly 9 9
value 0 ly 4 4 ly 5 5 ly 6 6

free te 7 7 te 8 8 te 9 9
equal te 1 1 1 te 1 1
equal te 1 2 2 te 2 2
equal te 1 3 3 te 3 3
value 1 te 4 4 te 5 5 te 6 6

start 0.5 all

out mi iter=2000 adm=off sc ec

Analysis of NCP Wave 5 Satisfaction group 3
Data ni=9 no=422 ma=cm
Km
1
0 1
0 0 1
0 0 0 1
0 0 0 0.258501202434074 1
0 0 0 0.399282194184471 0.261742272196564 1
0 0 0 0.83750618127169 0.31305796293061 0.467848647944525 1
0 0 0 0.299381554367431 0.866349245574491 0.322513726997231 0.390483441905876 1
0 0 0 0.385403873150572 0.27828332494125 0.858186779550935 0.495217066132862 0.362135519215521 1

me
0.0 0.0 0.0 6.95971563981043 5.63438256658596 7.67146282973621 7.27536231884058 6.11463414634146 7.8641975308642

sd
1.0 1.0 1.0 2.00137432939421 2.3232634017412 2.05219770077402 1.85073030258621 2.28827787560755 1.95590680913828

model ny=9 ne=9 nk=6 ly=fu,fi te=di,fi ps=in be=in ga=in ph=in
!assuming that split-ballot groups are randomized, we do not expect differences between groups, so they are set to be invariant, except for those parameters affected by the randomization of the methods.
value 1 ly 4 4 ly 5 5 ly 6 6 ly 7 7 ly 8 8 ly 9 9
value 0 ly 1 1 ly 2 2 ly 3 3

equal te 1 4 4 te 4 4
equal te 1 5 5 te 5 5
equal te 1 6 6 te 6 6
equal te 2 7 7 te 7 7
equal te 2 8 8 te 8 8
equal te 2 9 9 te 9 9
value 1 te 1 1 te 2 2 te 3 3

start 0.5 all

out mi iter=2000 adm=off sc ec

Appendix C: Summary of final model adjustments of 5th NCP wave MTMM analyses

Experiment	Model adjustments	df	$\chi 2$	JRule
Norwegian political satisfaction	Free PH_{65} (correlation between method 3 and method 2). Free PH_{55} (method 2 variance) only in Group 3.	37	87,6	7
Trust in Norwegian institutions	Free GA_{55} and GA_{14} (method effects) in all groups.	37	88,83	6

In the first experiment, looking at the observed correlations we identified differences between groups in method 2. Thus, in group 3 we allowed the variance of method 2 to be different from group 1. Moreover, JRule suggested allowing a correlation between method 2 and method 3 in group 3, which is a reasonable adjustment because these two methods are very similar. The number of possible misspecified parameters left, detected by JRule, was 7.

In relation to the second experiment, JRule suggested to free the method effect parameters GA_{14} and GA_{55}. That adjustment was in this case made in all groups. The number of possible misspecified parameters left, detected by JRule, was 6.