

Cernat, Alexandru; Oberski, Daniel

Research Report

SOEP-IS 2014 - Separating systematic measurement error components using MTMM in longitudinal studies

SOEP Survey Papers, No. 1102

Provided in Cooperation with:

German Institute for Economic Research (DIW Berlin)

Suggested Citation: Cernat, Alexandru; Oberski, Daniel (2022) : SOEP-IS 2014 - Separating systematic measurement error components using MTMM in longitudinal studies, SOEP Survey Papers, No. 1102, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:

<https://hdl.handle.net/10419/253397>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by-sa/4.0/>

1102₂₀₂₂

SOEP Survey Papers
Series H - SOEP-IS Modules

SOEP-IS 2014 – Separating systematic measurement error components using MTMM in longitudinal studies

Alexandru Cernat and Daniel Oberski

Running since 1984, the German Socio-Economic Panel (SOEP) is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research, DIW Berlin.

The aim of the SOEP Survey Papers Series is to thoroughly document the survey's data collection and data processing.

The SOEP Survey Papers is comprised of the following series:

Series A – Survey Instruments (Erhebungsinstrumente)

Series B – Survey Reports (Methodenberichte)

Series C – Data Documentation (Datendokumentationen)

Series D – Variable Descriptions and Coding

Series E – SOEPmonitors

Series F – SOEP Newsletters

Series G – General Issues and Teaching Materials

Series H – SOEP-IS Modules

The SOEP Survey Papers are available at <http://www.diw.de/soepsurveyspapers>

Editors:

Dr. Jan Goebel, DIW Berlin

Prof. Dr. Stefan Liebig, DIW Berlin and Freie Universität Berlin

Prof. Dr. David Richter, DIW Berlin and Freie Universität Berlin

Prof. Dr. Carsten Schröder, DIW Berlin and Freie Universität Berlin

Prof. Dr. Jürgen Schupp, DIW Berlin and Freie Universität Berlin

Prof. Dr. Sabine Zinn, DIW Berlin and Humboldt Universität zu Berlin

Please cite this paper as follows:

Alexandru Cernat and Daniel Oberski. 2022. SOEP-IS 2014 – Separating systematic measurement error components using MTMM in longitudinal studies. SOEP Survey Papers 1102: Series H. Berlin: DIW/SOEP



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

© 2022 by SOEP

ISSN: 2193-5580 (online)

DIW Berlin
German Socio-Economic Panel (SOEP)
Mohrenstr. 58
10117 Berlin
Germany

soeppapers@diw.de

SOEP-IS 2014 – Separating systematic measurement error components using MTMM in longitudinal studies

Module Title in SOEP Documentation: Separating Systematic Measurement Error Components Using MTMM in Longitudinal Studies

Alexandru Cernat and Daniel Oberski

Separating systematic measurement error components using MTMM in longitudinal studies

Alexandru Cernat, Institute for Social and Economic Research, University of Essex
(acerna@essex.ac.uk)

Daniel Oberski, Department of Methodology and Statistics, Tilburg University
(d.oberski@tilburguniversity.edu)

Research question and motivation

Measurement error is a pervasive problem in survey research and can cause results to be severely biased. As is well known, attenuation of relationship estimates such as correlations and subgroup differences occurs when non-substantive variations in respondents' answers are unrelated to other answers – i.e. when the errors are random (Lord & Novick, 1968). *Systematic* measurement error can also occur, however, and can both attenuate and artificially increase apparent relationships (Andrews, 1984; Saris & Gallhofer, 2007). Particularly:

1. When some respondents tend to agree to "agree-disagree" questions while others do not, answers to such questions across respondents will be related due to **acquiescence** (Billiet & McClendon, 2000; Krosnick, 1991; McClendon, 1991; Ray, 1979, 1983) ;
2. When some respondents are more sensitive to social norms than others, dependence may arise between answers that are perceived as **socially desirable** (Belli, Traugott, Young, & McGonagle, 1999; Ganster, Hennessey, & Luthans, 1983; Paulhus, 1984);
3. Not all respondents are equally diligent: to get through the survey interview more quickly, some may devise a **satisficing** strategy by considering only the **extreme** (highest or lowest) response options, while others consider only the **middle** few options (Krosnick, Narayan, & Smith, 1996; Krosnick, 1991). Again, respondents with the extreme or middle response style will tend to provide answers that are similar for reasons unrelated to the question of interest to the researcher (Baumgartner & Steenkamp, 2001, 2006).
4. More generally, a given question can be asked using different methods, and respondents' answers may be related for reasons entirely unconnected with the true variables of interest, but purely due to common **method** variance (Andrews, 1984; Saris & Gallhofer, 2007).

Past research has sometimes found large effects of these systematic components (e.g., McClendon, 1991). It is not clear, however, what the relative importance of each factor might be. This has direct implications for survey question design. For instance, a question designer may note that acquiescence may pose a problem for agree-disagree questions (Hui & Triandis, 1985; Krosnick, 1991) and therefore choose to ask questions directly instead. However, if the direct method were to engender more method effects (e.g., McClendon, 1991), social desirability variance, or extreme response styles, the results could well be even more biased than they would have been with agree-disagree questions. It is therefore essential to estimate not just one type of systematic error effects, but also any trade-offs that might exist with other types of errors.

Another open question in the literature is the degree to which different types of respondents' systematic answer tendencies might be stable over time. Billiet & Davidov (2008) discuss a known case of acquiescence stability. Just as in cross-sectional data, systematic response tendencies may

distort relationship estimates. In longitudinal research any stability in systematic errors will distort longitudinal relationship estimates such as growth model estimates or autoregressive coefficients.

Systematic errors can rarely be observed directly, but the degree to which they influence the results can sometimes be estimated by a combination of data collection designs and model assumptions. A large literature exists on the application of particular models to estimate each of the above four components, assuming that the other components do not exist. For instance, in factor models for response style, it is assumed that there is no residual dependence within different methods of asking the questions (Billiet & McClendon, 2000). At the same time, in multitrait-multimethod models it is generally assumed that the factors do not correlate (Saris & Gallhofer, 2007), so that there is no style factor.

Thus, model assumptions for estimating different systematic measurement error components mutually contradict each other. These assumptions are, however, necessary for the purpose of identification: separate identification of the influence of each of these systematic error components is not usually allowed for by the data collection design. Two natural questions arise from here: to what degree are the currently used models misspecified and how do applied results differ once all of these factors are accounted for simultaneously?

Even more importantly, however, we need to know the effect of systematic errors on the main purpose and advantage of GSOEP-IS: longitudinal analysis. It is well-known that measurement unreliability spuriously causes more apparent change over time than has truly occurred, distorting analyses (Kessler & Greenberg, 1981; Plewis, 1985). However, if systematic errors occur, and are stable over time, these will cause the opposite: apparent stability over time in the respondents' attitudes that is only due to their consistent tendency to answer the questions in certain ways (Billiet & Davidov, 2008). Therefore, if these factors are not accounted for, substantive analyses of longitudinal data such as GSOEP-IS will be invalidated. We therefore propose to collect the longitudinal data over several waves, so that we may answer the questions: "how stable are the different systematic error factors over time?" and "to what extent does this threaten longitudinal analyses?"

Another key issue in measurement is cross-cultural equivalence (Baumgartner & Steenkamp, 2006; Davidov, 2010; Jowell, Kaase, Fitzgerald, & Eva, 2007; Kankaraš, Moors, & Vermunt, 2010). We know that respondents in different countries, for example Germany and the UK, can react differently to the same questions (e.g., Davidov, 2010). We do not know why. This experiment will allow us to investigate one particular explanation for cross-cultural differences in answer patterns: that the degree of systematic errors differ. A unique feature of our experiment is that it would allow us to tease apart the different systematic error sources so that the precise source of the cross-cultural difference can be pinpointed.

In this document we put forward a data collection design that does allow separate identification of different systematic error components over time as well as their comparison across two cultural contexts. We propose to do this by repeating an extended MTMM design in the GSOEP-IS¹. The design is extended because it includes not only method factors but also acquiescence, social desirability and extreme response styles. The similarity between the designs and the periods of data collection with the Understanding Society Innovation Panel would provide a wealth of data that can be used cross-nationally and longitudinally both for methodological and substantial research.

¹ The experiment has been recently accepted for the Innovation Panel of Understanding Society (UKHLS) and data collection in the UK will start next year.

The experiment detailed below will allow researchers:

- To **disentangle** these four sources of systematic measurement error from each other for a particular set of questions;
- To measure the **stability over time** of these sources of systematic measurement error;
- To establish the **influence on analyses of longitudinal change** of systematic measurement error in a panel study;
- To estimate the overall **level** at which measurement error, random and systematic, influences answers to these often-used questions;
- To compare cross-culturally **relative size of systematic and random errors and true scores**;
- To compare cross-culturally **stability of systematic errors and true scores**;
- To compare cross-culturally **theoretical models that explain systematic errors**.

Design specification

We suggest inserting 6 questions about immigration attitudes in the beginning of the questionnaire already being fielded, and 6 questions on the same topic in another format at the end, after 20 minutes or more of other questions asked. We therefore propose to add 12 questions to each respondent's interview. These items can easily be incorporated into the larger questionnaire or other experiments.

We propose to include all adult sample of the GSOEP-IS in the experiment. The main mode of the study, CAPI, is suitable for our research question.

We propose that the experiment is repeated for at least two additional waves. The interval proposed is every year. This design will enable us to answer the longitudinal questions presented previously. We propose that for each wave we re-randomise the respondents to one of the experimental groups. This would enable us to have both respondents that have the same methods and respondents who have changed.

We expect an individual questionnaire time of around 4 minutes. We took the number of questions for each respondent (12) and divided by 4 as a rule of thumb, then added another minute just to be on the safe side.

In addition to the main questions we propose 5 other questions for the entire adult sample (if not already measured). We emphasize that these are optional questions and while they are not essential for the design they would improve the final outputs. Three of the questions measure cognitive ability, one measures working memory while one measures left-right placement. The items are presented in the Annex.

The proposed items have already been fielded in Germany as part of the European Social Survey, where they did not pose any documented problems.

The first and second question formats depend on the group to which the respondent has been assigned. We suggest assigning respondents randomly to one of 56 groups. The rationale for forming these groups and the question formats is given below.

Further details of design:

The design can be described as an extended split-ballot multitrait-multimethod (MTMM) design. In the classical MTMM design (Campbell & Fiske, 1959), multiple traits are evaluated by multiple methods. Andrews (1984) suggested identifying "traits" with "survey items" and "methods" with "item wordings". In what follows we will refer to the combination of a survey item and a particular wording as a "survey question".

Andrews' method has been applied extensively to survey research (Saris & Andrews, 1991; Saris & Gallhofer, 2007; Saris & Gallhofer, 2007) particularly in the European Social Survey (ESS) (Saris et al., 2012), and allows for the separation of "trait" (item) variance from random error and common method variance. It does not, however, allow for other systematic errors identified by the literature: acquiescence and social desirability variance. We therefore extend the MTMM approach to allow for the separation of these two additional sources of systematic error variance. The method by which we extend the MTMM is simplest to conceive of as an adjustment to a within-subjects design.

The MTMM design can be seen as a within-subjects factorial experiment. Subject-time units are assigned to treatments (questions), which result from crossing the "trait" and "method" factors. The outcome is the respondent's answer to each question. For instance, "traits" could be whether the respondent thinks that (1) Germany should allow more immigration and (2) Germany is made a worse place to live due to immigration. "Methods" could be asking these items with (A) 2-point, (B) 5-point, and (C) 11-point agree-disagree scales. By crossing Trait \times Method, six treatments (question wordings) are obtained, which are presented to the respondent: {1A, 2A, 1B, 2B, 1C, 2C}.

The random effects assumption means that not all stimuli need to be presented to each respondent, but, as it turns out, only 2 per respondent are enough (Saris et al., 2004). Instead of {1A, 2A, 1B, 2B, 1C, 2C} asked in succession, each respondent is asked either {1A, 2A, 1B, 2B}, {1A, 2A, 1C, 2C}, or {1B, 2B, 1C, 2C}. Preferably the order of methods is randomized (A, B, or C first). There should be at least 20 minutes of other questions in between the two methods. This greatly increases the plausibility of the assumption that respondents' answers to different questions are independent given trait and method effects (Saris & Van Meurs, 1991).

Keeping the experimental design perspective in mind, it is not difficult to see how the MTMM design may be extended to include other factors besides trait and method effects. In particular, we are interested in separating not only method variance from trait and residual variance, but also acquiescence and social desirability variance. We therefore suggest to manipulate the following experimental factors:

- Number of scale points (method): 2 point or 11 point scale
- Socially desirable direction: positively or negatively formulated item on immigration
- Acquiescence direction: Agree-disagree or Disagree-agree scale

This yields $2 \times 2 \times 2 = 8$ possible item wordings (treatments) for each of the items. The appendix gives the exact wordings and how the final questionnaire versions are constructed.

Although there are 8 treatments, structural equation models depend only on the pairwise covariances of the variables, so that only combinations of pairs of wordings are required. For each pair of questions, the first format is presented early in the questionnaire, and the second format at the end of the questionnaire, with at least 20 minutes of other questions in-between. In our case there are

$\binom{8}{2} = 28$ possible pairs of question formats. The order of presentation within pairs should be randomized.

Planned output:

Our own analysis of these data will focus initially on disentangling the above four sources of systematic error from each other. We plan two waves of output, one after the initial wave and one after the third wave:

The first outputs will include:

- A paper for a methodological audience (e.g. Survey Research Methods (SRM), Sociological Methods & Research (SMR), Sociological Methodology, or JRSSA) on the design of this experiment in GSOEP and Understanding Society and how it can be applied to other surveys for the purpose of separating different systematic error components;
- Cross-cultural comparison of style factors and other correlated error sources in attitudes to immigration: Germany vs. the UK (e.g., Cross-Cultural Psychology, European J Political Research);

The outputs after the third wave will include at least:

- True versus observed change in attitudes towards immigration in Germany (Social Science Res, Kölner Zeitschrift für Soziologie und Sozialpsychologie, Zeitschrift für Soziologie);
- A paper for a methodological audience on the stability of systematic factors over time and the country differences between them (e.g. SEM, SMR, SRM).

In addition, we plan to present these topics highlighting the use of the GSOEP IS in international conferences such as ESRA, AAPOR, ITSEW, or RC33.

APPENDIX

Sample size motivation:

We suggest splitting the entire sample into 56 groups randomly and presenting each group with a different version of the questionnaire.

Although there are 56 groups, the two orderings of each pair can be combined, so that each covariance between observed variables will be calculated from $n/28$ observations. With a 100% response rate this would imply $3000/28 = 108$ (assuming a conservative individual sample size of 1500) observations per covariance and $3000/4 = 750$ per variance. With 50% nonresponse it would be about 54 observations per covariance and 375 per variance.

This appears rather low, but it should be kept in mind that the observed correlations are to be projected into a much smaller parameter space based on a SEM model. A recommendation in the SEM literature is that there be at least between 10 and 20 observations per parameter. In the most complex version of the model including all traits, there are 48 loadings and trait variances, 15 trait covariances, 2 method variances, 1 acquiescence factor variance, and 1 social desirability factor variance, leading to $48+15+4 = 67$ parameters. Following the SEM literature recommendations, there should then be at least between 10 and 20 times more observations, i.e. between 670 and 1340 observations.

We have also conducted a simulation study using the SEM software Mplus 6.12 to investigate whether, with a 25% response rate, the precision, coverage, and power of the variance parameters of interest would still be adequate. The Table below shows that this is the case. We used the PATMISS option in Mplus to simulate the planned missingness pattern. Assuming 750 responses, reliability coefficients around 0.7, and method, social desirability, and acquiescence standardised effects around 0.3 (10% of total variance), the power to detect these factor variances is well over 0.9. The power for the social desirability factor is lowest, and drops to 0.77 when all factor variances are set to 5% instead of 10%. Using 250 replications the estimates are unbiased and the standard errors are acceptable. Incidentally, this analysis also shows that the model we propose to apply to the experimental data is identified. The Mplus input and output for this simulation can be obtained from <http://daob.nl/wp-content/uploads/2013/05/MTMM-simulation.zip>.

	True value	Avg. est	Sim. sd	Avg. se	MSE	Coverage	Power
Allow1	0.55	0.541	0.105	0.106	0.011	0.944	1.00
Allow2	0.55	0.553	0.101	0.109	0.010	0.948	1.00
Bad1	0.55	0.545	0.109	0.107	0.012	0.928	1.00
Bad2	0.55	0.534	0.114	0.107	0.013	0.880	1.00
Social							
desirability	0.1	0.106	0.032	0.028	0.001	0.916	0.94
Method 1	0.1	0.102	0.019	0.019	0.000	0.948	1.00
Method 2	0.1	0.097	0.014	0.015	0.000	0.952	1.00
Acquiescence	0.1	0.100	0.011	0.011	0.000	0.936	1.00

Proposed questions

TEMPLATE FOR THE QUESTIONNAIRE

Now some questions about people from other countries coming to live in Germany.

<First set of 6 questions>

<Other interview questions>

To help us improve our questions in the future, here are some final questions on a range of different topics which are similar to previous ones. Please don't try to remember what you answered before but treat them as if they were completely new questions.

Now some questions about people from other countries coming to live in Germany.

<second set of 6 questions>

For the 6 items, "positive" and "negative" formulations. "positive" means that with a disagree-agree scale, the socially desirable direction will be towards the higher end of the scale.

Trait number	SD direction	Item formulation
T1	Positive	Germany should allow more people of the same race or ethnic group as most German people to come and live here
T2	Positive	Germany should allow more people of a different race or ethnic group from most German people to come and live here
T3	Positive	Germany should allow more people from the poorer countries outside Europe to come and live here
T4	Positive	It is generally good for Germany's economy that people come to live here from other countries
T5	Positive	Germany's cultural life is generally enriched by people coming to live here from other countries
T6	Positive	Germany is made a better place to live by people coming to live here from other countries
T1	Negative	Germany should allow fewer people of the same race or ethnic group as most German people to come and live here
T2	Negative	Germany should allow fewer people of a different race or ethnic group from most German people to come and live here
T3	Negative	Germany should allow fewer people from the poorer countries outside Europe to come and live here
T4	Negative	It is generally bad for Germany's economy that people come to live here from other countries
T5	Negative	Germany's cultural life is generally undermined by people coming to live here from other countries
T6	Negative	Germany is made a worse place to live by people coming to live here from other countries

There are 8 different wordings of each item, corresponding to combinations of three factors: the higher- or lower-end being the socially desirable direction, the number of scale points, and whether agree-disagree or disagree-agree questions are used. These lead to 8 wordings W1-W8; an example formulation for trait one is given in the last column.

Wording number	Social desirability	Number of scale points	Agree or Disagree	Required direction	Item formulation (using trait 1 as an example)
W1	Higher	2	AD	Negative	Germany should allow fewer people of the same race or ethnic group as most German people to come and live here
W2	Lower	2	AD	Positive	Germany should allow more people of the same race or ethnic group as most German people to come and live here
W3	Higher	11	AD	Negative	Germany should allow fewer people of the same race or ethnic group as most German people to come and live here
W4	Lower	11	AD	Positive	Germany should allow more people of the same race or ethnic group as most German people to come and live here
W5	Higher	2	DA	Positive	Germany should allow more people of the same race or ethnic group as most German people to come and live here
W6	Lower	2	DA	Negative	Germany should allow fewer people of the same race or ethnic group as most German people to come and live here
W7	Higher	11	DA	Positive	Germany should allow more people of the same race or ethnic group as most German people to come and live here
W8	Lower	11	DA	Negative	Germany should allow fewer people of the same race or ethnic group as most German people to come and live here

Instead of presenting each respondent with all 8 different wordings of the same items, any one respondent need only answer 2 different wordings of the items, one at the beginning and the second at the end of the questionnaire. There are chosen $(8,2) = 28$ different questionnaire versions corresponding to the combinations of wordings. Moreover the ordering of the wordings should be randomized, leading to 56 different conditions to be randomized in total.

	Randomized ordering 1 2			Randomized ordering 2 1	
Version number	Beginning	End	Version number	Beginning	End
V1	W1	W2	V29	W2	W1
V2	W1	W3	V30	W3	W1
V3	W1	W4	V31	W4	W1
V4	W1	W5	V32	W5	W1
V5	W1	W6	V33	W6	W1
V6	W1	W7	V34	W7	W1
V7	W1	W8	V35	W8	W1
V8	W2	W3	V36	W3	W2
V9	W2	W4	V37	W4	W2
V10	W2	W5	V38	W5	W2
V11	W2	W6	V39	W6	W2
V12	W2	W7	V40	W7	W2
V13	W2	W8	V41	W8	W2
V14	W3	W4	V42	W4	W3
V15	W3	W5	V43	W5	W3
V16	W3	W6	V44	W6	W3
V17	W3	W7	V45	W7	W3
V18	W3	W8	V46	W8	W3
V19	W4	W5	V47	W5	W4
V20	W4	W6	V48	W6	W4
V21	W4	W7	V49	W7	W4
V22	W4	W8	V50	W8	W4
V23	W5	W6	V51	W6	W5
V24	W5	W7	V52	W7	W5
V25	W5	W8	V53	W8	W5
V26	W6	W7	V54	W7	W6
V27	W6	W8	V55	W8	W6
V28	W7	W8	V56	W8	W7

Optional Questions

Political orientation:

In politics people sometimes talk of left and right. Where would you place yourself on a scale from 0 to 10 where 0 means the left and 10 means the right?

Cognitive ability:

Introduction

Next I'm going to read you several numbers and I'd like you to write them down. There will be a blank number in the series that I read to you. I would like you to tell me what number goes in the blank. For example, if I said the numbers "1, 2, BLANK 4" then what number would go in the blank?

Question 1:

Let's try another one. I'm going to read you a series of numbers. There will be a blank number in the series that I read to you. I would like you to tell me what number goes in the blank. 2...4...6...Blank (Now look at the numbers that you just wrote down and tell me the number that belongs in the blank.)

1. Correct answer (8)
2. Incorrect answer

Question 2:

Please write down the following numbers: **23 . . . 26. . . 30 . . . 35 . . . BLANK.** (Now look at the numbers that you just wrote down and tell me the number that belongs in the blank.)

1. Correct answer (41)
2. Incorrect answer

Question 3:

Next, please write down the following numbers: **18 . . . 10 . . . 6 . . . BLANK . . . 3.** (Now look at the numbers that you just wrote down and tell me the number that belongs in the blank.)

1. Correct answer (4)
2. Incorrect answer

Working memory:

Backward digit span: Participants are presented with a series of digits (e.g., '8, 5, 4') and must immediately repeat them back **in reverse order**. If they do this successfully, they are given a longer list (e.g., '1, 4, 6, 5'). The length of the longest list a person can remember in reverse is that person's backward digit span.

Length of longest list remembered in reverse: ____ (0-9)

REFERENCES

- Andrews, F. M. (1984). Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach. *Public Opinion Quarterly*, 48(2), 409–442. doi:10.1086/268840
- Baumgartner, H., & Steenkamp, J. E. M. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research*, 38, 143–156.
- Baumgartner, H., & Steenkamp, J.-B. E. M. (2006). An Extended Paradigm for Measurement Analysis of Marketing Constructs Applicable to Panel Data. *Journal of Marketing Research (JMR)*, 43(3), 431–442.
- Belli, R. F., Traugott, M. W., Young, M., & McGonagle, K. A. (1999). Reducing Vote Overreporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring. *The Public Opinion Quarterly*, 63(1), 90–108.
- Billiet, J., & Davidov, E. (2008). Testing the Stability of an Acquiescence Style Factor Behind Two Interrelated Substantive Variables in a Panel Design. *Sociological Methods & Research*, 36(4), 542–562. doi:10.1177/0049124107313901
- Billiet, J., & McClendon, M. (2000). Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items. *Structural Equation Modeling: A Multidisciplinary Journal*, 7(4), 608–628. doi:10.1207/S15328007SEM0704_5
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105. doi:10.1037/h0046016
- Davidov, E. (2010). Testing for comparability of human values across countries and time with the third round of the European Social Survey. *International Journal of Comparative Sociology*, 51(3), 171–191. doi:10.1177/0020715210363534
- Ganster, D. C., Hennessey, H. W., & Luthans, F. (1983). Social Desirability Response Effects: Three Alternative Models. *The Academy of Management Journal*, 26(2), 321–331. doi:10.2307/255979
- Hui, C. H., & Triandis, H. C. (1985). The Instability of Response Sets. *Public Opinion Quarterly*, 49(2), 253–260. doi:10.1086/268918

- Jowell, R., Kaase, M., Fitzgerald, R., & Eva, G. (2007). The European Social Survey as a measurement model. In R. Jowell, C. Roberts, R. Fitzgerald, & G. Eva (Eds.), *Measuring attitudes cross-nationally lessons from the European Social Survey* (pp. 1–31). Los Angeles; London: SAGE.
- Kankaraš, M., Moors, G., & Vermunt, J. (2010). Testing for Measurement Invariance with Latent Class Analysis. In Eldad Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-Cultural Analysis: Methods and Applications* (pp. 359–384). Routledge Academic.
- Kessler, R. C., & Greenberg, D. F. (1981). *Linear panel analysis: models of quantitative change*. Academic Press.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology*, 5(3), 213–236.
- Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New directions for evaluation*, 1996(70), 29–44.
- Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley Publishing Company, Inc.
- McClendon, M. (1991). Acquiescence and Recency Response-Order Effects in Interview Surveys. *Sociological Methods & Research*, 20(1), 60–103. doi:10.1177/0049124191020001003
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of personality and social psychology*, 46(3), 598.
- Plewis, I. (1985). *Analysing change: measurement and explanation using longitudinal data*. J. Wiley.
- Ray, J. (1979). Is the Acquiescent Response Style Problem Not So Mythical After All? Some Results From a Successful Balanced F Scale. *Journal of Personality Assessment*, 43(6), 638–643. doi:10.1207/s15327752jpa4306_14
- Ray, J. (1983). Reviving the Problem of Acquiescent Response Bias. *The Journal of Social Psychology*, 121(1), 81–96. doi:10.1080/00224545.1983.9924470
- Saris, W. E., & Gallhofer, I. N. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research* (1st ed.). Wiley-Interscience.

- Saris, W., & Gallhofer, I. (2007). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey research methods*, 1(1), 29–43.
- Saris, W., Oberski, D., R´villa, M., Rojas, D. Z., Lilleoja, L. G. I., & Gruner, T. (2012). *Final report about the project JRA3 as part of ESS Infrastructure (SQP 2002-2011)*. European Social Survey. Retrieved from http://www.upf.edu/survey/_pdf/RECSM_wp024.pdf
- Saris, Willem, & Andrews, F. (1991). Evaluation of Measurement Instruments Using a Structural Modeling Approach. In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 575–597). New York: Wiley-Interscience Publication.
- Saris, Willem, & Meurs, A. V. (1991). *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait Multimethod Studies*. North-Holand.