24

Extending the Within-Persons Experimental Design: The Multitrait-Multierror (MTME) Approach

Alexandru Cernat[1] *and Daniel L. Oberski*[2]

[1] *Social Statistics Department, University of Manchester, Manchester M13 9PL, UK*
[2] *Department of Methodology & Statistics, Utrecht University, Utrecht, 3584 CH, the Netherlands*

24.1 Introduction

The typical aim of surveys is analytical as they are used to help investigate relationships between variables. To this end, methodologists have strived to estimate and minimize errors that can bias such estimates. Examples of possible measurement errors that could bias these results are random error, interviewer effects, response styles, and processing errors. Typically, these can also vary over respondents and are therefore stochastic, with the known effect of biasing estimates of relationships (Fuller 1987). For example, random errors, method variance, and interviewer effects may severely distort the analytic goals of the survey researcher (Beullens and Loosveldt 2014; Krosnick 1999; Saris and Gallhofer 2007). While research has been carried out separately on each of these different types of measurement errors, the need to estimate multiple stochastic errors concurrently has also been highlighted in the Total Survey Error (TSE) framework (Groves and Lyberg 2010).

If these biasing effects are to be evaluated, corrected for, or minimized by design, we first need to know how strong they are. Quantifying the extent to which such errors are present in survey answers is thus an essential prerequisite to attaining the analytic goals of surveys. Additionally, doing so concurrently for multiple error sources is paramount if we are to understand the mechanisms that cause them and the tradeoffs they imply.

Many approaches to estimating the extent of stochastic errors have been put forward in the literature. Suggested designs are “validation” data or “record check” studies (e.g. Katosh and Traugott 1981), interview–reinterview (test–retest) designs (Battese et al., 1976), as well as psychological scale evaluations, multitrait-multimethod (MTMM) (Saris and Gallhofer 2007), quasi-simplex (Alwin 2007), latent class (Biemer 2011), and other latent variable model approaches to estimate acquiescence, “yea-saying”, extreme response style, or social desirability variance (Billiet and Davidov 2008; Billiet and McClendon 2000; Moors 2003; Moors et al., 2014; Oberski et al., 2012). Of these, the record check design is the strongest, if it can be assumed that the “validation” data are completely error-free themselves. Unfortunately, however, validation data are difficult and sometimes impossible to obtain, and the assumption of no measurement error is often questionable (Ansolabehere and Hersh 2010; Groen 2012). Although these approaches may appear very different, they are similar in that they all require

Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment, First Edition.
Edited by Paul J. Lavrakas, Michael W. Traugott, Courtney Kennedy, Allyson L. Holbrook, Edith D. de Leeuw, and Brady T. West.

Companion Website: www.wiley.com/go/Lavrakas/survey-research

some form of repeated data collection. From the perspective taken in this chapter, all methods to estimate the extent of stochastic error are a form of within-person design.

The within-persons approach necessary to quantify stochastic errors has three important drawbacks. First, it tends to allow for only one source of stochastic error, such as random error (Alwin 2007; Battese et al., 1976; Biemer 2011; Katosh and Traugott 1981) or two sources, such as random error and acquiescence response style (Billiet and McClendon 2000). Each of these approaches is therefore designed to model only one or two forms of stochastic error, assuming that the other forms are absent. Second, the designs often lack randomization of the order of different measurements (Alwin 2011), which means that carryover (Myers 1972) or test effects (Campbell and Stanley 1963) might bias the estimates. Third, respondents may remember their answer at the previous occasion in test–retest type designs such as the interview–reinterview, quasi-simplex, and MTMM designs. Although Saris and Andrews (1991) have suggested that more than 20 minutes between forms removes this effect, Alwin (2011) questioned this conclusion.

Due to the difficulties of the within-persons designs outlined above, Krosnick (2011) suggested abandoning them altogether. Unfortunately, if one abandons within-persons designs, the only alternative for evaluating stochastic errors is relying on validation data, which as noted above are generally nonexistent. Between-persons experiments indeed do not suffer from the above problems, but neither are they informative enough to estimate the stochastic components of TSE. Therefore, the view taken in this chapter is that we should not do away with within-persons experiments but strive to improve them.

This chapter introduces a general framework that uses a within-persons experimental design but deals with two of the three problems highlighted above: the assumption of only one error type and the assumption of zero test effects. The "multitrait-multierror" (MTME) framework does this by applying a simple idea: extend the within-persons design to vary several error sources at a time and randomize methodological variation such as question order. This design then enables researchers to concurrently estimate multiple sources of stochastic measurement errors from the TSE framework, allowing for the improvement of question design and removal of biasing effects from analyses.

In the next section, we will present the MTME framework. We then give practical advice on how to design and implement such experiments in surveys and on how to estimate the effects of the experimental treatments. The approach is illustrated using the Understanding Society Innovation Panel in the United Kingdom (UKHLS-IP). Finally, we comment on future research needed to improve within-persons designs, including the remaining problem of memory effects.

24.2 The Multitrait-Multierror (MTME) Framework

Our framework starts with the observation that any type of repeated observation on the same respondent can be viewed as a within-persons design. We will therefore use the term "within-persons design" to mean any situation in which multiple measures of the same variable have been obtained on the same respondent; we will call such designs "experiments" when the precise method of measurement and its timing is under the control of the researcher. Preferably, but not necessarily, such choices undergo randomization.

For instance, a record check study that observes respondents' answers to the question "have you voted in the last Presidential election?" together with an official vote record is a within-persons design, but it is not an experiment because the researcher does not control data collection in the administrative record. On the other hand, a survey company that calls back some of its respondents after a face-to-face interview to reinterview them is performing an experiment in our terms, because the company could have chosen to phone first, or reinterview them face-to-face. A *randomized* within-persons experiment might have randomized these choices among subgroups of respondents.

24.2.1 A Simple but Defective Design: The Interview–Reinterview from the MTME Perspective

In the familiar notation of Shadish et al. (2002), a possible interview–reinterview design to estimate mode effects could be depicted as:

Condition 1:	X_{CAPI}	O	X_{CATI}	O
Condition 2:	X_{CAPI}	O		

In this notation, the “X’s” indicate a “treatment,” and the “O’s” an observation. The subscripts “CAPI” and “CATI” have been used to indicate that the treatments correspond to these data collection modes; other aspects of the treatments may also be relevant, however. Here, condition 1 is “exposed” to a CAPI interview and observed (i.e. interviewed). At a later date, this group of people is exposed to CATI and observed. In contrast, condition 2 has only the CAPI interview. To clarify these conditions, the data obtained from such a design can be coded in a design matrix, as shown in Table 24.1. In this design matrix, different columns encode factors that may be of interest, or that may cause methodological variation.

Table 24.1 represents the long form data where each row is a person/time record. As can be seen from the first column, some individuals have two rows, meaning that they were observed twice, while others have only one. This can be due to either being in condition 2 (who were not reinterviewed) or to being a nonrespondent in the second observation of condition 1. The second column shows an important component of TSE: random response error. This error varies every time we have a new measure. Thus, here we have two instances of random error (1 and 2) for each measurement made within a person. The third column recognizes that the number of visits may change over time. The “Repetition” column in Table 24.1 encodes the possibility of test effects, such as respondent fatigue leading to satisficing (Krosnick 2011). The data collection mode is known to have an effect on answers as well, so that the number of remembered visits may differ over the telephone versus face-to-face. The topic is a constant: all measures obtained are about the number of doctor visits, meaning we can only generalize the results to the respondents’ doctor visits, which is intended in this case.

Table 24.1 The standard interview–reinterview design in which random error, true change, repetition, and data collection mode have been confounded.

Person ID	Random error	Change	Repetition	Mode	Topic
1	1	No	No	CAPI	Doctor’s visits
1	2	Yes	Yes	CATI	Doctor’s visits
2	1	No	No	CAPI	Doctor’s visits
3	1	No	No	CAPI	Doctor’s visits
3	2	Yes	Yes	CATI	Doctor’s visits
4	1	No	No	CAPI	Doctor’s visits
…	…	…	…	…	…

It should be clear from the design matrix in Table 24.1 that the standard interview–reinterview design leaves a lot to be desired. Inadvertently or not, a number of factors have been confounded with one another, so that random error, change, test effects, and data collection mode all vary together and cannot be disentangled. Another key point to note is that any Topic × Person interaction can be interpreted as the person's score for that topic, i.e. as their "true score" or "trait." In the above design, the Topic × Person interactions are the same as the Person main effects since Topic is constant. In other designs, with multiple topics, this will not be the case.

When the survey goal is to estimate relationships among variables, these relationships are primarily biased by *variation across persons* in the methodological factors, i.e. in the Factor × Person interactions and their correlations (Carroll et al. 2006; Fuller 1987). The Error × Person variance is one such example. It is known as the variance of the random errors, the error variance, random error or unreliable variance, and it is well-known to cause bias when estimating relationships between variables. From this point, we will refer to it as random variance. For example, Alwin (2007) found random variance to be around 35% of the total variation in self-reported questions in a number of surveys. This can attenuate correlations between variables, but the bias in estimates of relationships can be in either direction when we conduct multivariate analyses such as multiple regression. So, if we are interested in the relationship between the number of doctor visits and another self-report, such as physical health, with 35% of its total variance arising from random variance, a true correlation of 0.8 would be estimated as 0.5 using the observed data. From an experimental design perspective (e.g. Cox and Reid 2000, ch. 6), the main issue is the confounding among the Topic × Person and Error × Person design columns, so that error and true variance cannot be separated without a full-rank design. A full-rank design in this case would randomize the respondents to all possible combinations of the levels of our design factors.

In short, if a design can be developed so that the methodological factors' interactions with the Person factor are estimable, then the amount of bias in relationship estimates caused by these factors can be estimated. One goal in the above example is therefore to estimate how strong the effect of random error is relative to the overall variation in doctor's visits: expressed as a proportion of variance, this corresponds to the "(un)reliability" of the question. After collecting the data, the following model might be fitted to the *j*th observation on the *i*th person to estimate this effect:

$$\begin{aligned} y_{ij} = {} & \beta_0 + \beta_1 Change_j + \beta_2 Error_j + \beta_3 Repetition_j + \beta_4 Mode_j + \beta_5 Person_i \\ & + \beta_{0,5,i}(Topic_j)(Person_i) + \beta_{1,5,i}(Error_j)(Person_i) \end{aligned}$$

where the coding of the categorical variables Time, Repetition, Mode, and Person has been omitted for clarity. In this case, the variables are dummies (two categories), but the model can be easily extended to variables with multiple categories. Since the Topic is constant across conditions, it has been absorbed into the other terms. Note that the usual residual error term has been replaced here by an Error × Person interaction. This is an equivalent formulation, i.e. $\beta_{1,5,\mathrm{i}}$ plays the role of a residual here. As mentioned previously, in the interview–reinterview design, Repetition, Error, and Mode are all confounded with one another, and the Person main effect is confounded with the Topic × Person effect. This shows that for the classic interview–reinterview design, the following assumptions are necessary to identify the Error × Person interaction:

- There are no test, repetition, or mode effects. This implies that the effects of $Change_j$, $Error_j$, and $Repetition_j$ are all 0 (i.e. $\beta_2 = \beta_3 = \beta_4 = 0$).
- There is no $Person_j$ main effect ("style factor") beyond the person's true opinion (i.e. $\beta_5 = 0$).

- There are no further interactions with $Person_j$, e.g. the effects of $Change_j \times Person_j$, $Mode_j \times Person_j$, etc. are zero, as are any higher-order interactions. This would not be true if some people get fatigued faster than others, for example.

This leads to the model:

$$y_{ij} = \beta_0 + \beta_1 Change_j + \beta_{0,5,i}(Topic_j)(Person_i) + \beta_{1,5,i}(Error_j)(Person_i)$$

Assuming that all factors have been coded to sum to zero ("effect-coding"), β_0 is the mean number of doctor visits over people and repetitions, β_1 is a deviation from that mean depending on the first or second occasion of asking the question, the $\beta_{0,5,i}$ are person scores on doctor visits, and the $\beta_{1,5,i}$ are random error scores. Finally, taking $Person_j$ to be a random factor, the model can be estimated using linear random-effects modeling or, equivalently, confirmatory factor analysis. Either of these techniques will allow estimation of $\text{var}(\beta_{1,5,i})$, the error variance of interest and the corresponding reliability, $1 - \frac{var(\beta_{1,5,i})}{var(y_i)}$.

To conclude this example, the classic interview–reinterview design can be seen as a within-persons design. However, for it to yield the error variance (reliability) of interest, a number of strong assumptions are necessary. Moreover, the discussion above has only factored in a few of the possible components of TSE, allowing for none of them besides random error in the model. This design, therefore, is unrealistic in light of the previous empirical findings in survey methodology and from the theoretical considerations of the TSE framework.

24.2.2 Designs Estimating Stochastic Survey Errors

With the limitations of the basic interview–reinterview approach now spelled out as a within-persons design, a partial solution also presents itself. In theory, if we can create a within-persons factorial experiment that varied within the design, we will be able to account for all of the effects assumed to be 0 so far. This suggests the following procedure:

1. Define the main types of stochastic (i.e. varying between people) measurement error sources whose influence is to be estimated;
2. Manipulate the survey questions to vary these error sources' influence;
3. Collect data using a random probability sample of persons; and
4. Estimate an appropriate model, taking *Person* to be a random factor.

In practice, this is possible for many factors, but not for all of them. In particular, any Person × Repetition interactions will remain confounded with random error, so this approach can solve all but the memory effect problem.

Examples of this approach are some MTMM designs (Andrews 1984; W. Saris and Gallhofer 2007). In these designs, the errors defined to be of interest are (i.) random error, (ii.) "method effects," defined as *Person* × "Question formulation" interaction effects, and sometimes also (iii.) order effects. In addition, the design varies the topic ("trait") of the question and allows for differences in errors over the different traits and question formulations. In this context, we would be able to disentangle the "method effect," the impact of the wording of the question or of the response scale, from the "trait." An MTMM design matrix is shown in Table 24.2.

As can be seen in Table 24.2, the MTMM design still confounds *Change, Repetition*, and *Method*. However, the MTMM design is a considerable improvement over the interview–reinterview design in other respects. First, the crossing with the *Topic* factor allows the *Error* × *Person* interaction to vary by *Topic* and *Repetition/Method*. In other words, it is no longer necessary to assume that the error variance is equal between the two time

Table 24.2 Typical survey MTMM design.

Person ID	Error × Topic	Change	Repetition	Mode	Topic
1	E_1	No	No	CAPI	Doctor's visits
1	E_2	Yes	Yes	CAPI	Doctor's visits
1	E_3	No	No	CAPI	Smoking behavior
1	E_4	Yes	Yes	CAPI	Smoking behavior
1	E_5	No	No	CAPI	General health
1	E_6	Yes	Yes	CAPI	General health
2	E_1	No	No	CAPI	Doctor's visits
…	…	…	…	…	…

points. Second, under the assumption that there are no *Change* × *Person* or *Repetition* × *Person* (e.g. memory) effects, the *Method* × *Person* interactions ("method effects") are now identifiable, allowing the researcher to study another source of stochastic TSE. Third, mode is now constant, so that this factor is no longer confounded with *Repetition/Change/Method*.

A possible linear model for the design in Table 24.2 is, given observation y_{ijt}, where *i* indexes the person, *j* the repetition, and *t* the topic:

$$y_{ijt} = \beta_{0,jt}(Method_j)(Topic_t) + \beta_{2,it}(Topic_t)(Person_i) + \beta_{3,ij}(Method_j)(Person_i) \\ + \beta_{4,ijt}(Error_j)(Topic_t)(Person_i)$$

Here $\beta_{0,\,jt}$ represents the expected mean for each question and method, the $\beta_{2,\,it}$ represents impact of the *Topic* on the expected score (i.e. "trait") while $\beta_{3,\,ij}$ is the impact of method on the answers (i.e. "method" effects). Finally, $\beta_{4,\,ijt}$ represents deviations from the expected score given the method and the topic and it can be considered an estimate of random error.

If we take *Person* to be a random factor, the first equation can be rewritten as a confirmatory factor analysis (CFA) model.

$y_{\mathrm{ijt}} = \lambda_{\mathrm{it}}\ \mathrm{T_j} + \lambda_{\mathrm{ij}}\ \mathrm{M_k} + \mathrm{E_{ijt}}$

Thus, $(Topic_t)(Person_i)$ can be estimated as a latent variable coded as $\mathrm{T_j}$ (from "trait"), while the $(Method_j)(Person_i)$ quantity can be estimated using another latent variable $\mathrm{M_k}$ (from "method"). Finally, the random error can be estimated for each repetition and topic. The variances of these factors represent the stochastic effects we want to estimate. In a simple CFA model that just estimates the stochastic error, we can ignore the mean structure in the data and $(Method_j)(Topic_t)$ will be fixed to 0. Of course these can also be estimated if needed. The coefficients that link $\mathrm{T_j}$ and $\mathrm{M_k}$ to the observed scores (y_{ijt}), the λ's, are also known as loadings in CFA. In order to estimate the stochastic errors, these loadings should be fixed based on the experimental design. Examples of how to fix these will be shown below.

To sum up the MTME perspective on MTMM experiments, some of the disadvantages of the standard interview–reinterview design could be solved by varying additional factors in the within-persons design: this is what leads to the MTMM design (Campbell and Fiske 1959). However, the MTMM design has its own shortcomings. It does not recognize the possibility of an overall acquiescence random effect, for example, or of social desirability variance. The MTME approach suggests that such issues can be accounted for by continuing the same line of thought: those factors that are unaccounted for should be varied in the within-persons design. The resulting data can then be analyzed using CFA-type models. The following sections explain this extended within-persons approach in more detail.

24.3 Designing the MTME Experiment

Designing, implementing, and analyzing data from an MTME experiment can be daunting. As such, we put forward a list of questions researchers and practitioners need to consider when using the MTME approach. This is divided into two stages: designing the experiment and estimating the statistical model.

There are five essential questions researchers need to answer when implementing the MTME design.

24.3.1 What are the main types of measurement errors that should be estimated?

Before planning the MTME design, researchers should decide what types of stochastic errors are known/expected to have an impact on the measures of interest. For some types of survey questions, we can expect small amounts of error. Examples of these are sociodemographic questions, such as sex or age or other factual information, such as number of household members. For some other types of questions, we can expect specific types of stochastic errors. For example, social desirability can have an important impact on sensitive topics such as sexual behavior or income (DeMaio 1984; Tourangeau et al., 2000). Other types of survey questions might be influenced by other biases such as acquiescence, method effects, or extreme response styles. The researchers have to decide which types of stochastic errors are the most important for the key measures of interest.

In order to illustrate these points, we will assume here that researchers are interested in estimating method effects, acquiescence, and random variance in their variables of interest.

24.3.2 How can the questions be manipulated in order to estimate these types of error?

Once the researcher has decided on the types of systematic errors of interest, they must consider if it is possible to manipulate the survey attributes, the questions, or the response categories in order to impact these stochastic errors. For example, MTMM models often manipulate the response scale (e.g. number of response categories, labeling of the categories, their numbering) in order to estimate method effects. Similarly, acquiescence can be manipulated by changing the ordering of the labels used for the response categories. For example, instead of using Agree–Disagree response categories, they can be reversed, leading to a Disagree–Agree formulation. Social desirability can be manipulated in a number of ways. For example, the mode of the questions can be changed, as some modes (e.g. self-administered ones) are better at minimizing this type of systematic error. This approach has the disadvantage of confounding mode and social desirability. Another way could be to change the question wording or present vignettes that imply what is the socially desirable answer. Thus, people could be primed with knowledge about what the majority supports or does.

Once the researchers decide on the types of systematic errors and how to manipulate them, they have to decide on the number of levels for each treatment. For example, if researchers want to estimate acquiescence and method effects, they have to decide on the level and the types of treatments they want to apply. As such, in the case of acquiescence, they can have two levels of the treatment: Disagree–Agree response categories and Agree–Disagree categories. Any "acquiescence" effect would then presumably positively bias the first form while having the reverse effect on the second form. An additional form for which acquiescence effects could be assumed zero, such as item-specific scales, is also possible (Saris et al., 2010), but not considered in this example. For the method effect, there are a number of options depending on the number of response categories, the amount of labeling, and the numbering of the categories. Let us

Table 24.3 Four question "forms" as a result of combining two "methods" and two response scale orders.

Form	Method	Acquiescence
F_1	5 point	Agree/disagree
F_2	5 point	Disagree/agree
F_3	10 point	Agree/disagree
F_4	10 point	Disagree/agree

assume researchers choose two types of methods: a fully labeled five-point scale and a ten-point scale with only the extreme categories labeled.

The combination of the two acquiescence manipulations and two methods leads to four different "forms" of the questions (Table 24.3). Implementing the forms in the split-ballot design (Saris and Gallhofer 2007; Saris et al., 2004) leads to four combinations of two $\left(\binom{4}{2} = 6\right)$ forms. This implies that six different pairs of forms must be administered to the respondents. Such a design can be implemented by giving a pair of forms to one of six randomized groups.

24.3.3 Is it possible to manipulate the form pair order?

After deciding on the types of errors to be estimated, the treatments, and the form pairs, the researchers must decide if the order of the forms will be randomized. The advantage of implementing such an approach is that it tackles some of the possible carryover effects that can appear due to the lack of independence of the two within-person measurements. On the other hand, this can increase the amount of groups to be created and analyzed. While this does not have an effect on respondent burden, as each individual still receives two measures, it does have an effect on the resources needed by the data collection agency and the analysts. If the researchers decide to randomize the order in our hypothetical example, this results in 12 form combinations (6 × 2). These are presented in Table 24.4.

In conclusion, to implement this MTME design that makes it possible to estimate method effects, acquiescence and random variance while controlling for order effects, the researcher must create 12 random groups, each of which will receive a combination of two formats of the questions.

24.3.4 Is there enough power to estimate the model?

When dividing the sample into such a large number of groups, the power of the analysis has to be taken into consideration. The first advantage of using the latent variable modeling framework

Table 24.4 Six form combinations with randomized order.

Number	Time 1	Time 2	Number	Time 1	Time 2
1	F_1	F_2	7	F_2	F_1
2	F_1	F_3	8	F_3	F_1
3	F_1	F_4	9	F_4	F_1
4	F_2	F_3	10	F_3	F_2
5	F_2	F_4	11	F_4	F_2
6	F_3	F_4	12	F_4	F_3

is the ease of implementing maximum likelihood methods for dealing with missing data. This approach uses all the available information in the analysis, maximizing power. Additionally, as the groups are randomized, no bias will be introduced as missing information is missing completely at random (see for an overview Enders 2010).

Nevertheless, even with this statistical method of dealing with missing data, enough information must be present to estimate each parameter. It is good practice to first implement a simulation study to consider the power of the design under different (conservative) nonresponse rates.

24.3.5 How can data collection minimize memory effects?

As mentioned in the previous section, memory effects (or other carryover effects) are an important threat to the validity of within-persons designs. That is also true for MTME experiments. Nevertheless, researchers can adopt a number of strategies in order to minimize the possibility of such bias.

One approach is to minimize memory effects by design. This can be done in a number of ways, for example, by having a minimum period of time between the two measurements, such as the 20 minutes proposed by Saris and Andrews (1991). If it is possible to collect data again, at a different point in time, researchers have to take into consideration two different aspects: memory and change. The ideal period for collecting the second measure in a within-persons experimental design is one that minimizes both any memory effects and change in the true score. The nature of these two dimensions depends on the topic of the questions used in the experiment. If the second point is in the same interview, then the distance should be maximized with the first measure being implemented as early as possible and the second one toward the end of the questionnaire.

Identifying the optimal time within the interviews to collect the second measurement is often challenging. To aid in this task, researchers can collect paradata to facilitate sensitivity analyses. One class of paradata that can be collected is time stamps or time latencies between the first measurement and the second one. These can now be easily collected in most computer-assisted data collection software. Similarly, researchers can collect information regarding individuals' memory capabilities. These two measures can be used after data collection for sensitivity analysis by estimating their effect on the MTME coefficients. It should be noted that these approaches are not ideal, as possible confounds exist in such observational designs. Nevertheless, such sensitivity analyses might prove useful and insightful by providing evidence regarding the presence or absence of memory effects.

24.4 Statistical Estimation for the MTME Approach

The statistical estimation of the MTME model is closely linked to the latent variable modeling tradition of MTMM (Andrews 1984; Campbell and Fiske 1959; Eid 2000; Saris et al., 2004). As such, each observed item is a combination of the true/trait score (*Person* × *Topic*) and stochastic error (*Person* × *Error* source). The contribution of the MTME approach is the possibility of experimentally manipulating multiple types of systematic errors concurrently while explicitly controlling for order effects.

In the MTME experiment proposed previously, we included both method and acquiescence as treatments. The model can be written as:

$$y_{\mathrm{jkl}} = \lambda_{\mathrm{Tjkl}}\mathrm{T}_{\mathrm{j}} + \lambda_{\mathrm{Mjkl}}\mathrm{M}_{\mathrm{k}} + \lambda_{\mathrm{Ajkl}}\mathrm{A}_{\mathrm{l}} + \mathrm{E}_{\mathrm{jkl}}$$

Where the observed items Y_{jkl} are measured by trait (question) j, method k with acquiescence effect l and decomposed into trait T_j, method M_k, acquiescence A_l and a specific residual

Table 24.5 Using the MTME design matrix to inform the model constraints needed for estimation.

Form	Method	Acquiescence
F_1	0	+1
F_2	0	−1
F_3	1	+1
F_4	1	−1

component, E_{jkl}. Additionally, the $\lambda\,T_{jkl}$ are the trait loadings, the $\lambda\,M_{jkl}$ are the method loadings, and the $\lambda\,A_{jkl}$ are the acquiescence loadings.

The design matrix (Table 24.5) can be used as a guide to the constraints needed for estimation. For the method effect, we use "dummy" (0/1) coding, which corresponds to the "MTM(M-1)" coding proposed for such models proposed by Eid (2000) and Eid et al. (2003). Thus, only one method is represented by a latent variable while the other is considered as a reference. Here we estimate the effect of method 2 (10-point scale) by fixing loadings of the items measured using forms 3 and 4 to +1. We estimate acquiescence as explained by Billiet and McClendon (2000) and Billiet and Davidov (2008), using one latent variable. The questions measured with forms 1 and 3 should have the loadings fixed to +1 as the Disagree-Agree response scales suffer this directional bias, while questions measured with forms 2 and 4 should have the loadings fixed to −1, as the Agree–Disagree wording is thought to reverse acquiescence effects relative to Disagree–Agree.

We can also use a graphical representation as another way to understand the statistical model needed to estimate the different types of errors. Figure 24.1 presents the equivalent of the formula above in the visual form. Here large circles represent latent variables/random effects, while squares represent observed variables. For example, the four squares are the same question, Q_1, measured using the 4 different forms, F_1–F_4. Each arrow can be conceived as a regression

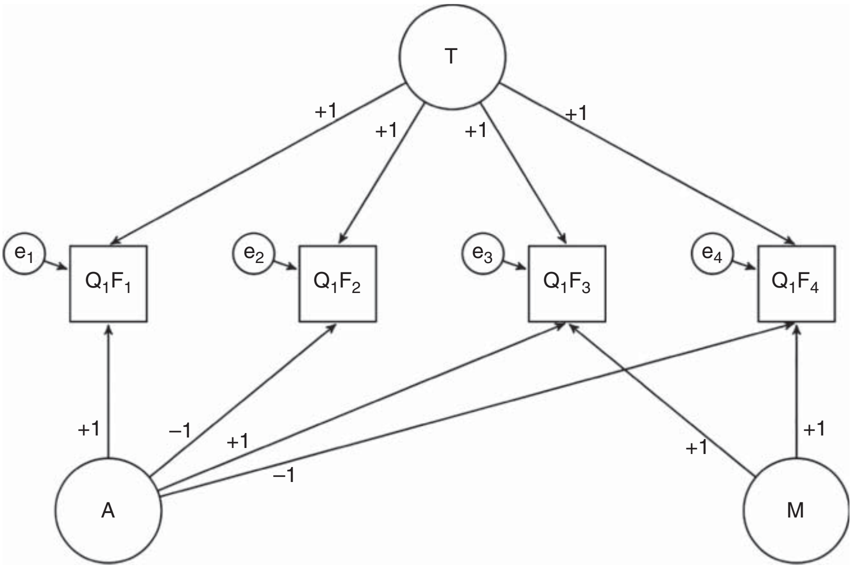

Figure 24.1 Visual representation of the Structural Equation Model for MTME with two stochastic errors.

coefficient. They can be either unrestricted, meaning they have to be estimated, or they can be restricted. For example, we can see that the arrows from A, the estimate of acquiescence, have been coded as shown in Table 24.5. Forms 1 and 3 should have higher acquiescence so we expect a positive relationship, while forms 2 and 4 should have less of it, thus we fix them to −1. We do the same for the Method effect, M. In the graphical representation of the SEM, any missing arrows between objects assume that there is no relationship, i.e. the coefficient is 0. Thus, the lack of arrows from M to the question measured in the first two forms is equivalent to having arrows restricted to 0, and is in accordance with Table 24.5. The small circles (coded e_1–e_4) are the residuals, i.e. what remains unexplained after we control for the other three sources of variance. It can be considered an estimate of unreliability/random error. The T in the figure represents an estimation of the "true" score or "trait" after controlling for the other influences. It is our estimate of what people would have answered to question 1 if we control for acquiescence (A), method effect (M), and random error (e_1–e_4).

24.5 Measurement Error in Attitudes Toward Migrants in the UK

In this section, we will give an example of an MTME experiment implemented in the UK Household Longitudinal Study – Innovation Panel (UKHLS-IP)[1]. The measures of interest are attitudes toward immigrants (Table 24.6) and have been previously used in other surveys such as the European Social Survey. In the first subsection, we will go through the five points discussed above in order to highlight the design process and how we implement it. In the second part, we will present the first results from the analysis. This is just one possible design. Researchers should adapt the MTME approach and analysis to best fit their needs.

24.5.1 Estimating Four Stochastic Error Variances Using MTME

24.5.1.1 What are the main types of measurement errors that should be estimated?

Attitudinal questions are notoriously hard to measure as they are less stable than values and much more subjective and prone to misunderstanding than factual questions. Because of their ephemeral nature, they can be easily influenced by response scale formatting. This influence

Table 24.6 Six questions ("traits") measuring attitudes toward immigrants in the UKHLS-IP.

Trait	Wording
Q_1	The UK should allow **more** people of the same race or ethnic group as most British people to come and live here
Q_2	UK should allow **more** people of a different race or ethnic group from most British people to come and live here
Q_3	UK should allow **more** people from the poorer countries outside Europe to come and live here
Q_4	It is generally **good** for UK's economy that people come to live here from other countries
Q_5	UK's cultural life is generally **enriched** by people coming to live here from other countries
Q_6	UK is made a **better** place to live by people coming to live here from other countries

[1] In this study, we ignore weights in order to facilitate the understanding of the model and results. The aim of the analysis here is for illustration and not finite population inferences regarding these relationships. Weights can be easily included in structural equation modeling (Oberski 2014).

can appear in multiple forms, from method effects to acquiescence or extreme response styles. Additionally, some topics can be considered sensitive, thus increasing threats to validity. As such, we believe that multiple sources of errors must be taken into consideration to validly measure some attitudinal scales.

Here we decided to manipulate two formatting characteristics that might bias answers to attitudinal questions: method and acquiescence. We also believe that questions regarding migration are prone to social desirability bias as there are important cultural norms and debates around this topic. As such, we also want to estimate the random effects of social desirability. Finally, random variance can also play an important role when collecting data about attitudes. This source of variation will be estimated by taking advantage of the within-persons nature of the MTME experimental design.

24.5.1.2 How can the questions be manipulated in order to estimate these types of errors?

To estimate the three types of systematic errors, two treatment levels were manipulated for each one:

- **Method**: Number of scale points (2 point vs. 11-point scale);
- **Acquiescence**: Agree–Disagree vs. Disagree–Agree response scale;
- **Social desirability**: positively vs. negatively formulated item on immigration.

This yields $2 \times 2 \times 2 = 8$ possible item wordings (forms) for each of the six items (traits). By combining them, there are ($\binom{8}{2} = 28$) possible pairs of question formats to be applied in the split-ballot MTME experiment. In this application, we randomized the order of the form pairs, thus leading to a total of 56 experimental groups (28×2). For example, we have implemented separately both a design that uses F_1 at the start of the survey and F_2 at the end as well as the reverse of this (F_2 at the start and F_1 at the end). This was done in order to minimize any potential carryover effects (Table 24.7).

24.5.1.3 Is it possible to manipulate the form pair order?

In this application, we have decided to randomize the order of the form pairs, thus leading to a total of 56 experimental groups (28×2). This was done in order to minimize any potential carryover effects.

24.5.1.4 Is there enough power to estimate the model?

While this design seems to define a large number of groups (i.e. small sample size per group), it should be kept in mind that the observed correlations are to be projected into a much smaller

Table 24.7 Eight forms to measure three types of systematic error in UKHLS-IP.

Form	Scale points	Agree–disagree	Social desirability	Wording
F_1	2	AD	Higher	Negative
F_2	2	AD	Lower	Positive
F_3	11	AD	Higher	Negative
F_4	11	AD	Lower	Positive
F_5	2	DA	Higher	Positive
F_6	2	DA	Lower	Negative
F_7	11	DA	Higher	Positive
F_8	11	DA	Lower	Negative

parameter space based on an SEM model. In the most complex version of our model, including all traits, there are 48 loadings and trait variances, 15 trait covariances, 1 method variance, 1 acquiescence factor variance, and 1 social desirability factor variance, leading to $48 + 15 + 3 = 66$ parameters.

We conducted a simulation study using the SEM software Mplus 6.12 to investigate whether, with a 50% response rate, the precision, coverage, and power of the variance parameters of interest would still be adequate. We used the PATMISS option in Mplus to simulate the planned missingness pattern. Assuming 750 responses, two traits,[2] reliability coefficients around 0.7, and method, social desirability, and acquiescence standardized 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. The Mplus syntax for the simulation is provided in the online appendix.

24.5.1.5 How can data collection minimize memory effects?

In order to minimize possible memory effects, the routing of the questionnaire avoided asking the second form if fewer than five minutes passed since the last question of the first form. Additionally, time stamps/latencies and cognitive ability were measured and will be used in the future for sensitivity analyses. This procedure would enable us to investigate some of the assumptions of the model that cannot be controlled for using the experimental design.

One possible way to use this information for sensitivity analysis after data collection could be to run the model separately for those that had the smallest difference between answering the two forms, which was five minutes, and those that had a longer period between them, for example, around 20 minutes. If memory effects are an issue, the correlation between the forms should be higher for the first group compared to the second one. Of course, this is not a perfect assessment of memory effects due to possible confounding factors. For example, some people might answer fewer questions due to routing in the questionnaire and their characteristics might be related to sources of error. Cognitive ability could provide a powerful control variable in this context.

24.5.2 Estimating MTME with four stochastic errors

As presented previously, the design matrix (Table 24.8) can be used to understand what latent variables must be estimated and what constraints have to be employed. In this case, we have six

Table 24.8 Design matrix for MTME model measuring attitudes toward immigrants in UKHLS-IP.

Form	Method	Acquiescence	Social desirability
F_1	0	+1	+1
F_2	0	+1	−1
F_3	1	+1	+1
F_4	1	+1	−1
F_5	0	−1	+1
F_6	0	−1	−1
F_7	1	−1	+1
F_8	1	−1	−1

[2] We use just two traits to make the simulation easier. In principle, the addition of traits should add more information to the model, making it easier to estimate.

trait measures of attitudes toward immigrants. Additionally, we have three types of systematic errors estimated as latent variables: method, acquiescence, and social desirability. The effect of the second method (11-point scale) is estimated by constraining all the items from forms 3, 4, 7, and 8 to +1. Acquiescence is measured as in the previous example with forms 1 to 6 having the loadings constrained to +1 (agree/disagree formats), while questions measured using forms 5 to 8 have the loadings constrained to −1. Using a similar approach, the social desirability latent variable model is estimated by constraining all the items in forms 1, 3, 5, and 7 to +1 and as −1 for the rest of them.

The relationships can also be written in equation form as shown in the formula below. Here we add the effect of social desirability, S_m, as measured by the λ_{Sjklm} loadings.

$$y_{jklm} + \lambda_{Tjklm}T_j + \lambda_{Mjklm}M_k + \lambda_{Ajklm}A_l + \lambda_{Sjklm}S_m + E_{klm}$$

The formula can also be illustrated in a SEM graphic (Figure 24.2). To keep things manageable, we have specified the model just for the first question/trait, but this can be easily extended to the other five. This time we have four sources of variation. In addition to acquiescence (A in the figure) and method (M) now, we also estimate social desirability (S), which we manipulated using the positive and negative wording of the questions. Once again we can use the design matrix in Table 24.8 to understand the restrictions in the coefficients. For example, questions asked in forms 1, 2, 5, and 6 were all measured using method 1 (2-point response scale). As such, they all have arrows restricted to 1 coming form M (which will represent method 1). Acquiescence and social desirability follow a similar pattern, having +1 for the forms where we expect a stronger effect and −1 when we expect a smaller one. We also see that T_1, the estimated "true" or "trait" score for question 1, has loadings fixed to 1 for all 8 questions (i.e. same question measured in eight different forms). In the case of our data, we would have five more questions, each measured using the eight forms. So, the model would be extended to include T2 to T6. The patterns of loadings would be similar to those shown in Figure 24.2.

24.6 Results

The design was implemented in the 7th wave of the UKHLS-IP (University of Essex. Institute for Social and Economic Research 2016). This is a longitudinal household survey in the United Kingdom used for methodological research. Wave 7 achieved a 54% household response rate (1505 households) and a 67% (2337 respondents) individual response rate. For more details regarding the data collection, see Al-Baghal et al. (2015).

Figure 24.3 presents the initial results from this MTME experiment. The analysis decomposes the total variance of the observed items into different components: trait, random error, method, acquiescence, and social desirability. This is done for each of the eight forms for five of the traits in the stacked bars as shown in Figure 24.3.

The total "quality" of the items can be described as those parts of the stacked bars that are trait variance (Saris and Gallhofer 2007). This quality varies considerably, from approximately 0.4 for form 8 and when asking "Allow different race" to approximately 0.9 for form 2.

With the current analysis, the highest quality was observed for question forms 1, 2, 5, and 6, which use the two-point response scale. It appears that their variance is less biased by the systematic errors included in this MTME design as compared to an 11-point answer scale.

These findings correspond to those of Revilla et al. (2014), who observed greater method variance for agree–disagree scales with 11 points than with fewer scale points, whereas they observed the converse with item-specific scales. Whether this finding can be generalized or is a consequence of our model formulation remains a topic for further investigation.

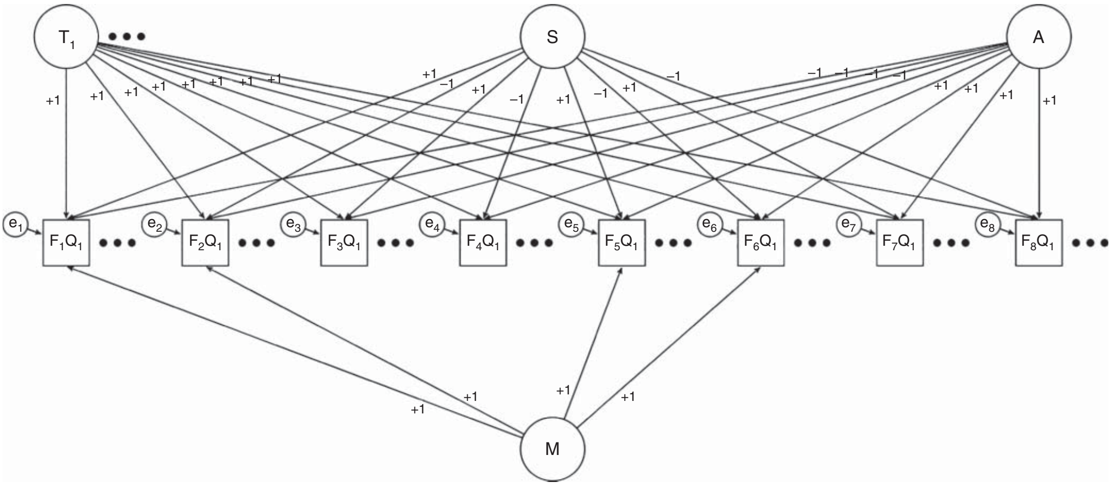

Figure 24.2 Visual representation of the Structural Equation Model for MTME with three stochastic errors. The analysis included six questions (Q_1–Q_6) regarding attitudes toward immigrants. Here only the model for Q_1 is shown for brevity (ellipses show how the model would expand for Q_2–Q_5).

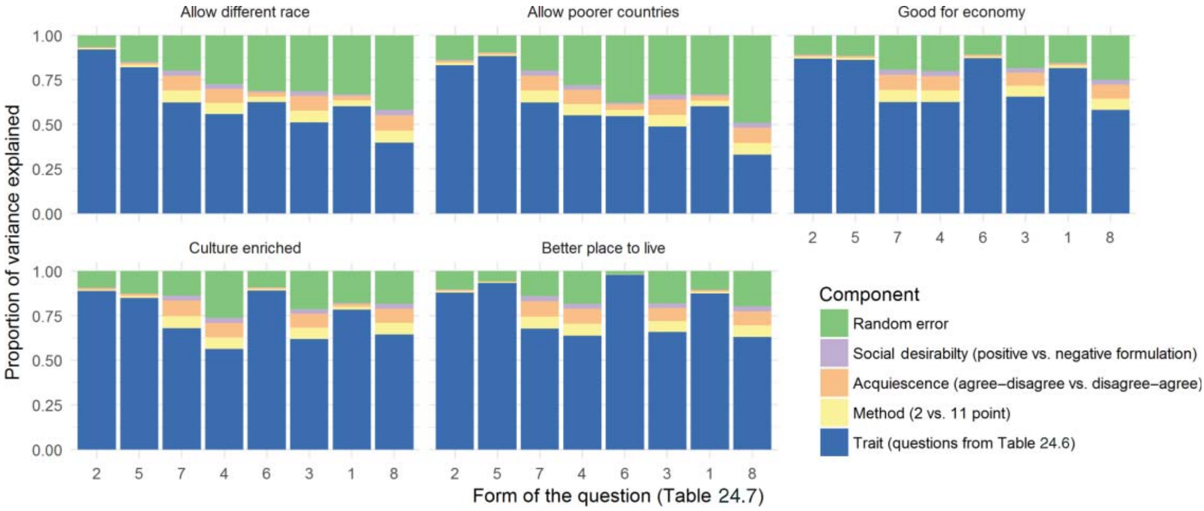

Figure 24.3 Variance decomposition in measures of attitudes toward immigrants.

Interestingly, the largest amount of nontrait variance is explained by random error in all the different questions and forms. This is followed by method and acquiescence, whereas social desirability does not explain much of the variance in the responses in our model.

24.7 Conclusions and Future Research Directions

In this chapter, we have argued that within-persons experimental designs are essential in survey research as they enable us to estimate and correct for stochastic errors that can bias substantive results. We have proposed a new design, the MTME design that directly tackles two of the problems with previous approaches. First, it enables researchers to concurrently estimate multiple types of systematic errors. Second, by randomizing the order of question forms, it makes it possible to control for some of the carryover or test effects.

We have encouraged the reader to answer five essential questions in order to design an MTME experiment:

1. *What are the main types of measurement errors that should be estimated?*
2. *How can the questions be manipulated in order to estimate these types of error?*
3. *Is it possible to manipulate the form pair order?*
4. *Is there enough power to estimate the models?*
5. *How can data collection minimize memory effects?*

We have also shown an application of the MTME in the UKHLS-IP, which is publicly available.[3] In this design, we have used six traits measuring attitudes toward immigrants and estimated concurrently four types of stochastic errors: method effects, acquiescence, social desirability, and random error. We have seen that around 70% of the variation comes from "trait," or is valid variation, while the rest is explained by the other components. While this is in line with some other research (e.g. Alwin 2007; Saris and Gallhofer 2007), there are very different estimates of data quality depending on the wording of the questions. Again, this is in line with the large body of research in survey methodology that has tried to develop a set of best practices for wording and formatting questions. Interestingly, our results indicate that random error represents the biggest proportion of nontrait variance. This is unexpected given the sensitive nature of the questions, in which we expect higher social desirability bias.

This application is just an example of the possible ways in which survey questions can be manipulated and stochastic error variances estimated using MTME. An area ripe for future development is how to extend this approach and think of new and creative ways to manipulate questions in order to estimate stochastic variance.

There is one important assumption of within-persons experiments that the MTME does not tackle: the memory effect. In design terms, we are left with a potential confounding of Person × Repetition with Person × Topic interactions. If these are present, they could bias the results from MTME experiments. Creative thinking is also needed here. In this chapter, we have proposed using paradata to see how the amount of time between the two measures or individual memory capabilities influence the model coefficients. This information should come at relatively low cost to the data collection agency. If possible, design solutions should be implemented to solve this issue. For example, in a web survey environment, it would be relatively easy to have the second reinterview at a later date, thus minimizing memory effects. If such an approach is used, two new aspects must be considered. First, the second measurement must be chosen such as to minimize both memory effects and changes in the true score. Second, the additional wave of data collection might increase the chances of nonresponse.

[3] https://www.understandingsociety.ac.uk/about/innovation-panel

In summary, since stochastic errors are an unavoidable part of surveys and we are often interested in studying relationships, within-persons designs are indispensible. The MTME approach goes some of the way toward clarifying how such designs help and how various sources of methodological bias can be mitigated while studying several sources of survey error simultaneously. At the same time, we recognize the potential problem of memory effects and suggest that future research programs should focus on reducing this concern with within-persons designs.

Acknowledgments

This paper makes use of data from the Understanding Society Innovation Panel. Understanding Society is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The research data are distributed by the UK Data Service.

References

Al-Baghal, T., Bloom, A., Burton, J. et al. (2015). Understanding society innovation panel wave 7: results from methodological experiments. *Understanding Society Working Paper Series* 2015–3: 1–62.

Alwin, D. (2007). *Margins of Error: A Study of Reliability in Survey Measurement*. New York: Wiley-Interscience.

Alwin, D. (2011). Evaluating the reliability and validity of survey interview data using the MTMM approach. In: *Question Evaluation Methods: Contributing to the Science of Data Quality*, 1e (ed. J. Madans, K. Miller, A. Maitland and G. Willis), 265–294. Hoboken, NJ: Wiley.

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

Ansolabehere, S. and Hersh, E. (2010). *The Quality of Voter Registration Records: A State-by-State Analysis*. Cambridge, MA: Department of Government, Harvard University https://elections.wi.gov/node/1234.

Battese, G., Fuller, W., and Hickman, R. (1976). Estimation of response variances from interview reinterview surveys. *Journal of the Indian Society of Agricultural Statistics* 28: 1–14.

Beullens, K. and Loosveldt, G. (2014). Interviewer effects on latent constructs in survey research. *Journal of Survey Statistics and Methodology* 2 (4): 433–458. https://doi.org/10.1093/jssam/smu019.

Biemer, P. (2011). *Latent Class Analysis of Survey Error*. New York: Wiley.

Billiet, J. and Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods and Research* 36 (4): 542–562. https://doi.org/10.1177/0049124107313901.

Billiet, J. and McClendon, M. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling: A Multidisciplinary Journal* 7 (4): 608–628. https://doi.org/10.1207/S15328007SEM0704_5.

Campbell, D.T. and Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56 (2): 81–105. https://doi.org/10.1037/h0046016.

Campbell, D.T. and Stanley, J.C. (1963). *Experimental and Quasi-Experimental Research*. Houghton Mifflin Company.

Carroll, R.J., Ruppert, D., Stefanski, L.A., and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC Press.

Cox, D.R. and Reid, N. (2000). *The Theory of the Design of Experiments*. CRC Press.

DeMaio, T. (1984). Social desirability and survey measurement: a review. In: *Surveying Subjective Phenomena* (ed. C. Turner and E. Martin), 257–282. New York: Russell Sage Foundation.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika* 65 (2): 241–261. https://doi.org/10.1007/BF02294377.

Eid, M., Lischetzke, T., Nussbeck, F.W., and Trierweiler, L.I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: a multiple-indicator CT-C(M-1) model. *Psychological Methods* 8 (1): 38–60. https://doi.org/10.1037/1082-989X.8.1.38.

Enders, C.K. (2010). *Applied Missing Data Analysis*, 1e. New York: The Guilford Press.

Fuller, W.A. (1987). *Measurement Error Models*. New York: Wiley.

Groen, J.A. (2012). Sources of error in survey and administrative data: the importance of reporting procedures. *Journal of Official Statistics (JOS)* 28 (2): 173–198.

Groves, R.M. and Lyberg, L. (2010). Total survey error: past, present, and future. *Public Opinion Quarterly* 74 (5): 849–879. https://doi.org/10.1093/poq/nfq065.

Katosh, J.P. and Traugott, M.W. (1981). The consequences of validated and self-reported voting measures. *Public Opinion Quarterly* 45 (4): 519–535.

Krosnick, J.A. (1999). Survey research. *Annual Review of Psychology* 50 (1): 537–567. https://doi.org/10.1146/annurev.psych.50.1.537.

Krosnick, J.A. (2011). Experiments for evaluating survey questions. In: *Question Evaluation Methods: Contributing to the Science of Data Quality*, 1e (ed. J. Madans, K. Miller, A. Maitland and G. Willis), 265–294. Hoboken, NJ: Wiley.

Moors, G. (2003). Diagnosing response style behavior by means of a latent-class factor approach. Socio-demographic correlates of gender role attitudes and perceptions of ethnic discrimination. *Quality and Quantity* 37: 277–302.

Moors, G., Kieruj, N.D., and Vermunt, J.K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology* 44 (1): 369–399. https://doi.org/10.1177/0081175013516114.

Myers, J.L. (1972). *Fundamentals of Experimental Design*, 2e. Boston: Allyn and Bacon.

Oberski, D., Weber, W., and Révilla, M. (2012). The effect of individual characteristics on reports of social desirable attitudes towards immigration. In: *Methods, Theories, and Empirical Applications in the Social Sciences: Festschrift for Peter Schmidt*, 2012e (ed. S. Salzborn, E. Davidov and J. Reinecke), 151–158. VS Verlag für Sozialwissenschaften.

Oberski, D.L. (2014). lavaan.survey: an R package for complex survey analysis of structural equation models. *Journal of Statistical Software* 57 (1): 1–27. https://doi.org/10.18637/jss.v057.i01.

Revilla, M.A., Saris, W.E., and Krosnick, J.A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods and Research* 43 (1): 73–97. https://doi.org/10.1177/0049124113509605.

Saris, W. and Andrews, F. (1991). Evaluation of measurement instruments using a structural modeling approach. In: *Measurement Errors in Surveys* (ed. P. Biemer, R. Groves, L. Lyberg, et al.), 575–597. New York: Wiley-Interscience Publication.

Saris, W. and 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., Révilla, M., Krosnick, J.A., and 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., Satorra, A., and Coenders, G. (2004). A new approach to evaluating the quality of measurement instruments: the split-ballot MTMM design. *Sociological Methodology* 34 (1): 311–347.

Shadish, W., Cook, T., and Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont, CA: Wadsworth Cengage Learning.

Tourangeau, R., Rips, L.J., and Rasinski, K. (2000). *The Psychology of Survey Response*, 1e. Cambridge University Press.

University of Essex. Institute for Social and Economic Research (2016). *Understanding Society: Innovation Panel, Waves 1–8, 2008–2015*, 7e. UK Data Service. SN: 6849.