Extending the within-persons experimental design: The multitrait-multierror (MTME) approach

Alexandru Cernat

University of Manchester, United Kingdom

Daniel L. Oberski

Tilburg University, The Netherlands

January 15, 2016

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 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 will severely distort the analytic goals of the survey researcher (Krosnick, 1999; Saris and Gallhofer, 2007; Beullens and Loosveldt, 2014). 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 framework (TSE; Groves and Lyberg, 2010). Therefore, 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 (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 (e.g. Billiet and McClendon, 2000; Billiet and Davidov, 2008; Moors, 2003; Oberski et al., 2012; Moors et al., 2014). Of these, the record check design is 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 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 (Battese et al., 1976; Katosh and Traugott, 1981; Alwin, 2007; Biemer, 2011) 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 carry-over (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 Van Meurs (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, this would also entail abandoning any analytic goal for a survey, since, absent perfect validation data, stochastic errors can only be evaluated using within-persons designs. Between-persons experiments indeed do not suffer the above problems, but neither are they informative enough to estimate the stochastic components of Total Survey Error. 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 them. The approach is illustrated using the Understanding Society Innovation Panel in UK (UKHLS-IP). Finally, we comment on future research needed to improve within-persons designs, including the remaining problem of memory effects.

2 The multitrait-multierror (MTME) framework

Our framework starts from 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 has 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' answer 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.

2.1 A simple but defective design: the interview-reinterview from the MTME perspective

In the familiar notation of Shadish et al. (2002), the standard interview-reinterview design 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. To clarify these, the data obtained from such a design can be coded in a design matrix, as shown in Table 1. In this design matrix, different columns encode factors that may be of interest, or that may cause methodological variation.

Table 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 re-interviewed) or to being a non-respondent in the second observation of condition 1. The second column shows an important component of Total Survey Error: random response error. This error will vary between the two occasions. The third column recognizes that the number of visits may change over time. The "Repetition" column in Table 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's visits, meaning we can only generalize the

Table 1: The standard interview-reinterview design. 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
...	...	...	...	...	...

results to the respondents' doctor's visits, which is intended in this case.

It should be clear from the design matrix in Table 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 pried apart. Another key point to note is that any Topic $\times$ Person interaction can be interpreted as the person's score for that topic, i.e. as their "true score" or "trait". While in the above design, the Topic $\times$ 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 $\times$ Person interactions and their correlations (Fuller, 1987; Carroll et al., 2006). For example, the Error $\times$ Person variance, known as the "error variance" in the measurement error literature, is well-known to cause more bias in relationship estimates as it increases. From an experimental design perspective (e.g. Cox and Reid, 2000, ch. 6), this can be understood as a confounding among the Topic $\times$ Person and Error $\times$ Person design columns, so that error and true variance cannot be separated without a full-rank design.

In short, if a design can be found 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 applied to the j-th observation on the i-th person to estimate this:

$$y_{ij} = \beta_0 + \beta_1 \text{Change}_j + \beta_2 \text{Error}_j + \beta_3 \text{Repetition}_j + \beta_4 \text{Mode}_j + \beta_5 \text{Person}_i + \beta_{0,5,i}(\text{Topic}_j)(\text{Person}_i) + \beta_{1,5,i}(\text{Error}_j)(\text{Person}_i), \quad (1)$$

where the coding of the categorical variables Time, Repetition, Mode, and Person has been omitted for clarity. 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 a Error $\times$ Person interaction. This is an equivalent formulation–i.e. $\beta_{1,5,i}$ 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 $\times$ Person effect. This shows that for the classic interview-reinterview design, the following assumptions are necessary to identify the $\beta_{1,5,i}$ effects:

- There are no test, repetition, or mode effects, i.e. $\beta_2 = \beta_3 = \beta_4 = 0$;
- There is no Person main effect ("style factor") beyond the person's true opinion, i.e. $\beta_5 = 0$;
- There are no further interactions with Person, e.g. the effects of Change $\times$ Person, Mode $\times$ Person etc. are zero, as are any higher-order interactions.

This leads to

$$y_{ij} = \beta_0 + \beta_1 \text{Change}_j + \beta_{0,5,i}(\text{Topic}_j)(\text{Person}_i) + \beta_{1,5,i}(\text{Error}_j)(\text{Person}_i). \quad (2)$$

Assuming all factors have been coded to sum to zero ("effects-coding"), β_0 is the mean number of doctors' 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 doctors' visits, and the $\beta_{1,5,i}$ are random error scores. Finally, taking Person 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, $1 - \frac{\text{var}(\beta_{1,5,i})}{\text{var}(y_i)}$.

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 substantive interest, a number of heroic assumptions are necessary. Moreover, the discussion above has only factored in a few of the possible components of total survey error, 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 total survey error framework.

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 such that the TSE error sources assumed zero have been varied within the design, we will be able to account for all of these

Table 2: Typical survey MTMM design.

Person ID	Error × Topic	Change	Repetition	Mode	Method	Topic
1	E1	No	No	CAPI	M1	Doctor's visits
1	E2	Yes	Yes	CAPI	M2	Doctor's visits
1	E3	No	No	CAPI	M1	Smoking behavior
1	E4	Yes	Yes	CAPI	M2	Smoking behavior
1	E5	No	No	CAPI	M1	General health
1	E6	Yes	Yes	CAPI	M2	General health
2	E1	No	No	CAPI	M1	Doctor's visits
...	...	...	...	...	...	...

effects. This suggests the following procedure:

1. Define the main types of stochastic 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;

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 that this approach can solve all but the memory effect problem.

Examples of this approach are some "multitrait-multimethod" (MTMM) designs (Andrews, 1984; Saris and Gallhofer, 2007). In these, 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. An MTMM design matrix is shown in Table 2.

As can be seen in Table 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 $\times$ 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 points. Second, under the assumption that there are no Change $\times$ Person or Repetition $\times$ Person (e.g. memory) effects, the Method $\times$ 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 2 is, given observation y_{ijt}, where i indexes the person, j the repetition, and t the topic,

$$\begin{aligned} y_{ijt} = & \\ & \tau^*_{jt}(\text{Method}_j)(\text{Topic}_t)+ \\ & \eta^*_{ijt}(\text{Topic}_j)(\text{Person}_i)+ \qquad (3) \\ & \xi^*_{ij}(\text{Method}_j)(\text{Person}_i)+ \\ & \epsilon^*_{ijt}(\text{Error}_j)(\text{Topic}_t)(\text{Person}_i) \\ & \qquad = \tau_{jt} + \lambda_{jt}\eta_i + \xi_i + \epsilon_{ijt}. \qquad (4) \end{aligned}$$

Again, if we take Person to be a random factor, Equation 4 can be recognized as a confirmatory factor analysis model. An adjustment to the standard hierarchical linear model has been made here, by allowing the random Person $\times$ Topic effects to be multiplied by a factor ("loading") λ_{jt} for each within-person observation. Generally these loadings are relative to the first within-person observation for that topic, i.e. $\lambda_{1t} := 1$.

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 multitrait multimethod 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 confirmatory factor-type models. The following sections explain this extended within-persons approach in more detail.

3 Designing the MTME

Designing, implementing and analyzing data from an multitrait multierror 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 in two stages: designing the experiment and estimating the statistical model.

We will now discuss five essential questions researchers need to answer when implementing the multitrait-multierror design.

1. *What are the main types of measurement errors that should be estimated?*

   Before planning the MTME design, researchers have to 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 socio-demographic 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 behaviors or income (Tourangeau et al., 2000; DeMaio, 1984). Other types of survey questions might be influenced by other biases such as acquiescence, methods effects or extreme response style. 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, acquiescence and random variance in their variables of interest.

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 presenting vignettes that imply what is the social 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 ones. 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 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 3). Implementing the forms in the split-ballot design (Saris et al., 2004; Saris and Gallhofer, 2007) leads to four combinations of two ($\binom{4}{2} = 6$) 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.

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 random-

Table 3: Four questions "forms" as a result of combining two "methods" and two response scale orders

Form	**Method**	**Acquiescence**
F1	5 point	Agree/Disagree
F2	5 point	Disagree/Agree
F3	10 point	Agree/Disagree
F4	10 point	Disagree/Agree

Table 4: Six form combinations with randomized order

Nr.	**Time 1**	**Time 2**	**Nr.**	**Time 1**	**Time 2**
1	F1	F2	7	F2	F1
2	F1	F3	8	F3	F1
3	F1	F4	9	F4	F1
4	F2	F3	10	F3	F2
5	F2	F4	11	F4	F2
6	F3	F4	12	F4	F3

ized. The advantage of implementing such an approach is that it tackles some of the possible carry-over 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 has one on the resources needed by the data collection agency and the analysts. If the researchers decide to randomized the order in our hypothetical example this results in 12 form combinations (6×2). These are presented in Table 4.

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

4. *Is there enough power to estimate the model?*

   When dividing the sample in 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 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 Enders, 2010, for an overview).

   Nevertheless, even with this statistical method of dealing with missing data enough information must be present to estimate each parameter. We believe it to be a good practice to implement a simulation study to consider the power of the design under different (conservative) non-response rates.

5. *How can data collection minimize memory effects?*

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

   One approach to the issue of memory effects is to minimize it by design. This can be done in a number of ways, for example by having a minimum criterion of time between the two measures, such as the 20 minutes proposed by Saris and Van Meurs (1991). If there is the flexibility of collecting 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 towards the end of the questionnaire.

Considering the typical difficulty in choosing the time for the second measurement researchers can collect paradata to facilitate sensitivity analyses. A first type of measures that can be collected are 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.

4 Statistical estimation of the MTME

The statistical estimation of the MTME is closely linked to the latent variable modeling tradition of MTMM (Campbell and Fiske, 1959; Andrews, 1984; Saris and Andrews, 1991; Saris et al., 2004; Eid, 2000). As such, each observed item is a combination of the true score (Person $\times$ Topic) and stochastic error (Person $\times$ Error source). The contribution of the MTME is the possibility of manipulating multiple types of systematic errors concurrently while explicitly controlling for order effects.

In the MTME experiment proposed previously, we included both method and

Table 5: Using the design matrix for model constrains and estimation

Form	**Method$_2$**	**Acquiescence**
F1	0	+1
F2	0	-1
F3	1	+1
F4	1	-1

acquiescence as treatments. The model can be written as:

$$Y_{jkl} = \lambda_{Tjkl} T_j + \lambda_{Mjkl} M_k + \lambda_{Ajkl} A_l + E_{jkl} \tag{5}$$

Where the observed items Y_{jkl} are measured by trait (question) j, method k with acquiescence effect l and is decomposed in trait T_j, method M_k, acquiescence A_l and specific residual component, E_{jkl}. Additionally, λ_{Tjkl} are the trait loadings, λ_{Mjkl} are the method loadings and λ_{Ajkl} are the acquiescence loadings.

The design matrix (Table 5) can be used as a guide to the necessary 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); 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 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.

Table 6: Six questions ("traits") measuring attitudes towards immigrants

Trait	Wording
T1	The UK should allow **more** people of the same race or ethnic group as most British people to come and live here
T2	UK should allow **more** people of a different race or ethnic group from most British people to come and live here
T3	UK should allow **more** people from the poorer countries outside Europe to come and live here
T4	It is generally **good** for UK?s economy that people come to live here from other countries
T5	UK?s cultural life is generally **enriched** by people coming to live here from other countries
T6	UK is made a **better** place to live by people coming to live here from other countries

5 Measurement error in attitudes towards migrants in UK

In this section we will give an example of a MTME experiment implemented in the UK Household Longitudinal Study - Innovation Panel (UKHLS-IP). The measures of interest are attitudes towards immigrants (Table 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 and researchers should adapt the MTME approach and analysis to best fit their needs.

5.1 Estimating four stochastic error variances using MTME

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 can appear in multiple forms, from methods effects to acquiescence or extreme response style. Additionally, some topics can be considered sensitive, thus increasing the threat 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 error can also play an important role when collecting data about attitudes. This will be estimated by taking advantage of the within persons nature of the MTME experimental design.

2. *How can the questions be manipulated in order to estimate these types of error?*

   To estimate the three types of systematic errors two treatments level 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

Table 7: Eight forms to measure three types of systematic error

Form	**Scale points**	**Agree-Disagree**	**Social desirability**	**Wording**
F1	2	AD	Higher	Negative
F2	2	AD	Lower	Positive
F3	11	AD	Higher	Negative
F4	11	AD	Lower	Positive
F5	2	DA	Higher	Positive
F6	2	DA	Lower	Negative
F7	11	DA	Higher	Positive
F8	11	DA	Lower	Negative

items. By combining them there are $\binom{8}{2} = 28$ possible pairs of question formats to be applied in the split-ballot MTME experiment.

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 carry-over effects.

4. *Is there enough power to estimate the model?*

   While these appear like 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 parameter space based on a 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 have 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, 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 syntax for the simulation is provided in the online appendix.

5. *How can data collection minimize memory effects?*

   In order to minimize possible memory effects the rooting of the questionnaire avoided asking the second form if fewer than 5 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.

5.2 Estimating MTME with four stochastic errors

As presented previously, the design matrix (Table 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 trait measures. Additionally, we have three types of systematic errors estimated as latent variables. 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

Table 8: Design matrix for MTME model measuring attitudes towards immigrants in UKHLS-IP

Form	**Method$_2$**	**Acquiescence**	**Desirability**
F1	0	+1	+1
F2	0	+1	-1
F3	1	+1	+1
F4	1	+1	-1
F5	0	-1	+1
F6	0	-1	-1
F7	1	-1	+1
F8	1	-1	-1

and as -1 for the rest of them.

The relationships can also be written in equation form as we have seen in formula 5. Here we add the effect of social desirability, S_m, as measured by the λ_{Sjklm} loadings.

$$Y_{jklm} = \lambda_{Tjklm}T_j + \lambda_{Mjklm}M_k + \lambda_{Ajklm}A_l + \lambda_{Sjklm}S_m + E_{jklm} \tag{6}$$

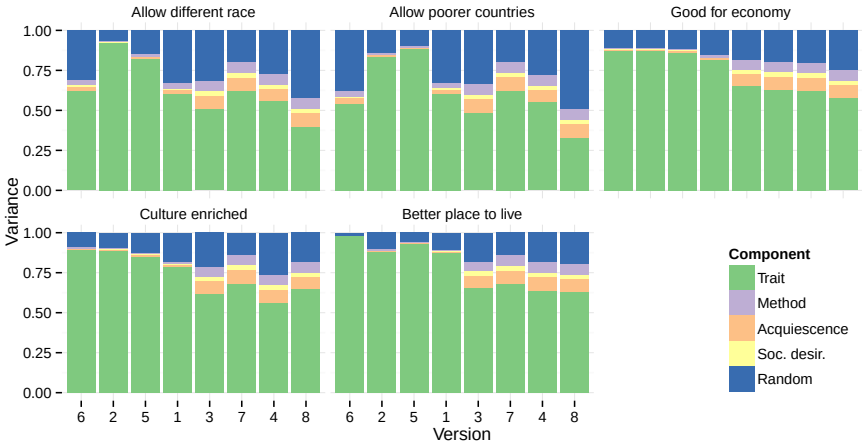
5.3 Implementation and results

The design was implemented in the 7^{th} wave of the UKHLS-IP. 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 66.7% (2337 respondents) individual response rate. For more details regarding data collection see Al-Baghal et al. (2015).

Figure 1 describes the initial results from this MTME experiment. The analysis decomposes the total variance of the observed items in the different components: trait, random error, method, acquiescence and social desirability. This is done for each of the 8 forms for five of the traits in the stacked bars shown in Figure 1.

The total "quality" of the items can be described as those parts of the stacked

Figure 1: Variance decomposition of attitudes towards immigrants



bars that are not trait variance (Saris and Gallhofer, 2007). This quality varies considerably, from approximately 0.5 for form 5 and trait 2 ("Allow different race") to approximately 0.9 for others. 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 may correspond to those of Révilla 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. Interestingly, the largest amount of non-trait 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.

6 Conclusions and future research program

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 multitrait multierror (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 questions forms, it makes it possible to control for some of the carry-over 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[1]. In this design we have used six traits measuring attitudes towards immigrants and estimated concurrently four types of stochastic errors: method effects, acquiescence, social desirability and random error. Results indicate random error represents the biggest proportion of non-trait variance.

This application is just an example of the possible ways in which questions in surveys can be manipulated and stochastic error variances estimated using MTME.

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

An area ripe for future development is how to extend this approach and thinking 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 non-response.

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 towards clarifying how such designs help, and how various sources of methodological bias can be mitigated while studying several TSE sources simultaneously. At the same time, we recognize the potential problem of memory effects and suggest future research programs should focus on reducing this concern with within-persons designs.

References

Al-Baghal, T., Bloom, A., Burton, J., Booker, C., Cernat, A., Fairbrother, M., Jackle, A., Kaminska, O., Keusch, F., Krosnick, J. A., Lynn, P., Oberski, D., Pudney, S., Sala, E., Schnettler, S., Silber, H., Stark, T., Uhrig, N., and Yan, T. (2015). Understanding Society Innovation Panel Wave 7: Results from Methodological Experiments. *Understanding Society Working Paper Series*, (2015-03):1–62.

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

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

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, 48(2):409–442.

Ansolabehere, S. and Hersh, E. (2010). The quality of voter registration records: A state-by-state analysis. *Cambridge, Mass.: Department of Government, Harvard University*.

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

Beullens, K. and Loosveldt, G. (2014). Interviewer effects on latent constructs in survey research. *Journal of Survey Statistics and Methodology*, page smu019.

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

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 & Research*, 36(4):542–562.

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.

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

Campbell, D. T. and Stanley, J. (1963). *Experimental and Quasi-Experimental Designs for Research*. Wadsworth Publishing, 1 edition.

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 Turner, C. and Martin, E., editors, *Surveying subjective phenomena*, pages 257–282. Russell Sage Foundation, New York.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, 65(2):241–261.

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.

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

Fuller, W. (1987). *Measurement Error Models*. John Wiley & Sons, New York.

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

Groves, R. M. and Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, 74(5):849–879.

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.

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

Moors, G. (2003). Diagnosing Response Style Behavior by Means of a Latent-Class Factor Approach. Socio-Demographic Correlates of Gender Role Attitudes and Percpetions of Ethnic Discrimination. *Quality & 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.

Myers, J. (1972). *Fundamentals of experimental design*. Allyn and Bacon, Boston, 2d ed. edition.

Oberski, D., Weber, W., and Révilla, M. (2012). The effect of individual characteristics on reports of socially desirable attitudes toward immigration. In Salzborn, S.,

Davidov, E., and Reinecke, J., editors, *Methods, Theories, and Empirical Applications in the Social Sciences: Festschrift for Peter Schmidt*. VS Verlag für Sozialwissenschaften.

Révilla, M. A., Saris, W. E., and Krosnick, J. A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods & Research*, 43(1):73–97.

Saris, W. and Andrews, F. (1991). Evaluation of measurement instruments using a structural modeling approach. In Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S., editors, *Measurement Errors in Surveys*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics Section, pages 575–597. Wiley-Interscience Publication, New York.

Saris, W. and Gallhofer, I. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Wiley-Interscience, 1 edition.

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.

Saris, W. and Van Meurs, A. (1991). *Evaluation of Measurement Instruments by Meta-Analysis of Multitrait Multimethod Studies*. North-Holand.

Saris, W. E., Révilla, M., Krosnick, J. A., and Shaeffer, E. M. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. In *Survey Research Methods*, volume 4, pages 61–79.

Shadish, W., Cook, T., and Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage Learning, Belmont, Calif.

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