Received: 17 September 2018 | Accepted: 30 May 2021

DOI: 10.1111/rssa.12733

ORIGINAL ARTICLE

Estimating stochastic survey response errors using the multitrait-multierror model

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

[1]Social Statistics, University of Manchester, Manchester, UK

[2]Department of Methodology and Statistics, University of Utrecht, Utrecht, The Netherlands

Correspondence
Alexandru Cernat, Social Statistics, University of Manchester, Manchester, M13 9PL, UK.
Email: alexandru.cernat@manchester.ac.uk

Funding information
ESRC National Centre for Research Methods, University of Southampton, Grant/Award Number: R121711

Abstract

Surveys are well known to contain response errors of different types, including acquiescence, social desirability, common method variance and random error simultaneously. Nevertheless, a single error source at a time is all that most methods developed to estimate and correct for such errors consider in practice. Consequently, estimation of response errors is inefficient, their relative importance is unknown and the optimal question format may not be discoverable. To remedy this situation, we demonstrate how multiple types of errors can be estimated concurrently with the recently introduced 'multitrait-multierror' (MTME) approach. MTME combines the theory of design of experiments with latent variable modelling to estimate response error variances of different error types simultaneously. This allows researchers to evaluate which errors are most impactful, and aids in the discovery of optimal question formats. We apply this approach using representative data from the United Kingdom to six survey items measuring attitudes towards immigrants that are commonly used across public opinion studies.

KEYWORDS

attitudes towards immigrants, design of experiments (DoE), latent variable modelling, measurement error, multitrait

1 | INTRODUCTION

Survey questions remain the primary instrument that pollsters use to tap into public opinion, governments use to count their citizens and social scientists use to study thoughts, feelings and behaviour. Questions, however, are subject to *response errors* (Alwin, 2007; Alwin & Krosnick, 1991; Saris & Gallhofer, 2007): systematic and random deviations of recorded answers from the truth. Response errors occur when any stage of the survey response process (Tourangeau et al., 2000) is 'satisficed' rather than 'optimized' by the respondent (Krosnick, 1991). Such errors can take on a variety of forms; we will discuss three here that are commonly studied. First, *acquiescence* refers to a tendency to agree with any statement, regardless of its content (Eckman et al., 2014; Krosnick & Presser, 2010). For example, a respondent may agree with an anti-immigration statement, but also with its opposite. Second, *social desirability* refers to a tendency to align answers with a perceived social norm, even when the true opinion differs from that norm (Fisher & Katz, 2000; Kreuter et al., 2008; Smith, 1967). For example, respondents who live in a left-leaning community may publicly declare more favourable attitudes towards immigration than they hold in private. Third, *method effect* ('common method variance') is a tendency to answer all questions asked in a specific manner in a similar direction (Campbell & Fiske, 1959; Saris & Gallhofer, 2007; Widaman, 1985). For example, a respondent may tend to choose the second answer to any two-point scale regardless of its labels ('recency' effect, Krosnick & Presser, 2010).

Response errors cause biases that can affect both the observed means and the (co)variances (Groves & Lyberg, 2010). Mean bias—the difference of the observed mean relative to a hypothetical mean free of response errors—occurs, for example, when acquiescence causes more 'agree' answers on average, when socially desirable answers are more popular, or when common method bias is directional, as is the case with recency effects in two-point scales. Another example could be the difference in average alcohol consumption between self-administered and interviewer surveys due to social desirability. While mean bias is commonly recognized, interest in social-scientific studies often focuses, not only on means, but also on variance and covariance structure: correlations, regression coefficients and multivariate analyses such as factor analysis or graphical models, for instance. Contrary to the intuition of some observers (e.g. Mayer-Schönberger & Cukier, 2014, ch. 3), such analyses can be shown to be very strongly affected by response error as well, under realistic circumstances.

(Co)variance bias will occur whenever respondents *differ* in their tendencies towards acquiescence, social desirability or method effect. For example, if perceived norms of what is socially desirable differ across people, answer tendencies will likewise differ across people. Such variation in response tendencies then causes additional variation in answers that is unrelated to the hypothetical true opinion of interest (Saris, 1988). Saris and Gallhofer (2007) distinguish between two types of such variation: random and correlated error. Random errors are individual differences in response error specific to a *single* question. Such errors are well known to bias regression coefficients (Fuller, 1987) and produce false appearances of change over time in longitudinal data (Cernat & Sakshaug, in press; Hagenaars, 2018). Correlated errors (Andrews, 1984), meanwhile, are respondent answer tendencies that apply across *different* survey questions and will therefore cause spurious correlations among the responses: the false appearance of relationships. This distorts multivariate statistical analyses (e.g. Pankowska et al., 2018; Spector et al., 2019) and leads to artificial stability over time in longitudinal data (Hagenaars, 2018). Because social-scientific studies often focus on (co)variance structure rather than means, here we will likewise focus on the factors that bias such analyses: random and correlated response errors.

Existing approaches to estimate random and correlated response errors, discussed below, include the 'quasi-simplex' and 'multitrait-multimethod' (MTMM) approaches (Alwin, 2007; Campbell & Fiske, 1959; Saris & Gallhofer, 2007). These do not distinguish different types of response errors, but instead focus on 'one factor at a time', an approach known as 'OFAT' in the design of experiments (DoE) literature (see e.g. Czitrom,1999 for why DoE has long since abandoned OFAT). As we explain below, OFAT is problematic because it (1) is a highly inefficient method of estimating the effects of response errors, (2) precludes direct comparison of the relative importance of different error sources and (3) yields suboptimal advice regarding question design.

In this paper, we propose a method of estimating the variance of different response errors simultaneously. We do so by combining fractional factorial *within-person designs* (Cox & Reid, 2000) with *latent variable models* (Oberski et al., 2017; Skrondal & Rabe-Hesketh, 2004). Our combined data collection-data analysis approach has an alternative interpretation as an extension of the MTMM approach—one that evaluates several response errors simultaneously; we therefore refer to it as the 'multitrait-multierror' (MTME) approach.

In the following section, we first discuss the history of response error variance estimation and the broad approaches that are encountered in the literature. Of particular interest is the classical MTMM approach which can be seen as a special case of MTME. Subsequently, we discuss the application to immigration attitude items, as well as the MTME approach to estimating variance due to acquiescence, social desirability and response scale ('method'). We then present the results and discuss some of the implications.

2 | BACKGROUND

Investigations of response bias have a rich history. Mean bias has received the most attention (e.g. Schuman & Presser, 1981); overviews are found in Schaeffer and Presser (2003), Krosnick and Presser (2010), and Oberski (2015). Random and correlated measurement error have generated a more modest body of evidence (for a review, see DeCastellarnau, 2017).

Traditionally, researchers have estimated mean bias by performing 'split-ballot' experiments: survey respondents are randomly assigned to one of (usually) two groups, and each group receives a different question format. For example, Schwarz et al., (1985) performed a highly cited experiment, which was later replicated extensively in the preregistered 'many labs' collaboration (Klein et al., 2014). The authors demonstrated that respondents claim to watch less television when the answer scale went up to 'more than two and a half hours' than if the scale went up to 'four and a half hours'. Thus, the response error source investigated by Schwarz and his colleagues is the value of the implicit norm communicated by the answer scale.

More recently, some researchers have managed to obtain a 'gold standard' record, that can be contrasted with the survey answer (e.g. Kreuter et al., 2010; Sinibaldi et al., 2013). The advantage of obtaining gold standard records is that very few assumptions are necessary to estimate the amount of error inherent in survey answers. The degree of error can be correlated with observational properties of the data collection, such as the interviewer (Sinibaldi et al., 2013), or true values (Kreuter et al., 2010). The disadvantages are that there are very few true gold standard values, since register data often contain considerable measurement error as well (Boeschoten et al., 2017; Oberski et al., 2017; Pankowska et al., 2018), and that it is not possible to observe a record of 'true opinion'. For these reasons, gold standard data, while providing a solution to the problem of both mean and (co)variance bias estimation in principle, are not often applicable in practice.

To estimate the effect of response errors on random and correlated measurement error without a gold standard, researchers often rely on a form of latent variable modelling (LVM, Bollen, 1989). The underlying idea of this approach is simple: repeated measures of the same unobserved ('latent') variable will share a common component. Identifying the functional relationship between this common component and the observed measures will then identify the 'quality' of these measures, that is, the extent of random and correlated error. The LVM approach is powerful, because it allows the researcher to estimate the extent of random and correlated error in a measure, without requiring gold standard variables. However, the LVM approach also has two requirements. First, it requires observing variables that can be seen as 'repeated measures of the same unobserved variable'. In other words, it requires a certain data collection *design*. Second, identification inevitably requires a set of assumptions regarding conditional independence of the observed variables. In other words, it relies on a *model*.

Here we discuss three broad approaches resulting from different design-model combinations in the LVM approach: the *internal consistency* approach, the *quasi-simplex* (or longitudinal) approach, and the MTMM approach. Each of these approaches, if deemed acceptable, can yield estimates of random and correlated error in survey items; each has, moreover, been used to relate such estimates to sources of response error.

The *internal consistency* consists of formulating survey questions that cover different topics which are assumed to form a 'scale' (Nunally & Bernstein, 1994). For example, a set of questions on the respondent's attitudes towards immigration might be surmised to measure a common underlying concept that could plausibly be called 'immigration attitude' (see Roots et al., 2016 and the first three questions in Table 1 for an example). Collecting answers to these questions (design) and fitting a confirmatory factor model to them (model) will yield factor loadings that, under the assumptions of this model, are proportional to the degree of random measurement error. If a second set of such items is brought into the analysis, it is possible to fit a bifactor model and estimate common 'style factors' (Billiet & McClendon, 2000), which are proportional to correlated errors.

The consistency approach only yields estimates *proportional* to random and correlated errors, because, in addition to response errors and common concept variance, each item will measure true opinions apart from this concept. For this reason, the loadings do not represent the reliability of each item, but rather its 'consistency' as a measure of the overall concept 'immigration attitude'

TABLE 1 The six questions used to measure attitudes towards immigration (original wording)

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

(Saris & Gallhofer, 2007). Consequently, the approach, even under its own assumptions, does not provide a direct estimate of random or correlated error. For this reason, Saris & Gallhofer (2007) and Alwin (2007) dismissed this as an approach to estimating response error variance. Moreover, the consistency model assumptions are readily violated when different items share a common true opinion that they do not share with the other items ('multidimensionality'), or exert a direct causal influence on the latent variable (bidirectionality) or on other observed variables.

A second, more targeted, approach to estimation of response error variance without a gold standard is the 'quasi-simplex' (longitudinal) approach (Alwin, 2007). In this approach, instead of three different items, the same item is repeatedly presented to the same respondent over a larger timespan, such as once per month or once per year. This design is already inherent to most panel surveys; thus, the quasi-simplex approach can usually operate by simply analysing existing data sources. Alwin (2007) suggested to apply the 'quasi-simplex' model to such data (Heise, 1969; Wiley & Wiley, 1974). The key assumption of this model is that there is no correlated error whatsoever, yielding conditional independence over time given the hidden states (latent variables). In addition, assumptions regarding the latent (hidden) transitions over time, and regarding the equality of error variances over time are commonly made. It was later shown that these latter two assumptions can often be relaxed while retaining identifiability of the key parameters (Jöreskog, 1978).

The quasi-simplex approach was critiqued by Saris & Gallhofer (2007, pp. 182–183). First, the quasi-simplex model considers time-specific variance as measurement error, which Saris and Gallhofer argued is inappropriate for opinion items. Second, Saris & Gallhofer argued that correlated error is often stable over time, and may therefore violate the core assumption of error independence across time. Violating this assumption can bias the comparison of survey items, because some apparently beneficial choices could have in fact have resulted from a larger degree of correlated error, causing detrimental effects to appear beneficial. In addition, assuming correlated error away, as this approach does, prevents us from studying how it is affected by response error sources.

Finally, Saris and Gallhofer (2007) and Saris et al., (2011) applied the 'multitrait-multimethod' (MTMM) approach. In this approach, data are collected cross-sectionally. At least three different survey items ('traits') are formulated using at least two different question formats ('methods'). Subsequently, the resulting six survey questions are presented to respondents in two blocks: one using method 1 for all three items towards the beginning of the questionnaire, and one using method 2 for the same three items, towards the end of the questionnaire. Saris & Van Meurs (1991) suggested at least 20 min of other questions should be presented to the respondent in between the two repetitions.

The MTMM *design* was analysed by these authors using the MTMM *model*. Let the centred observed answer to question form (method) m of survey item (trait) t be y_{tm}. Then the MTMM model is the linear confirmatory factor analysis model

$$y_{tm} = \lambda_{tm}^{(T)} T_t + \lambda_{tm}^{(M)} M_m + \varepsilon_{tm},$$

where traits T_t, methods M_m, and random errors ε_{tm} are random variables, all random variables are centred, and we assume independence between traits T_t, methods M_m, and random errors ε_{tm}, as well as among methods, that is, $E\left(T_t M_m\right) = E\left(M_{m'} M_m\right) = E\left(\varepsilon_{tm} M_m\right) = E\left(\varepsilon_{tm} T_t\right) = 0$, for all $t,\ m \neq m'$. Generally, for identification purposes the first indicator's trait loading is fixed to unity, $\lambda_{1m}^{(T)} = 1$, and method loadings are often set equal across traits, $\lambda_{tm}^{(M)} = 1$. Key parameters of interest are the estimated explained variance in y_{tm} due to the trait T_t ('reliability', complement of random

error) and due to the correlated error (method) factor M_m ('common method variance'). Extensions to categorical and other types of data and arbitrary distributions on the latent variables have also been introduced (Oberski, 2015; Oberski et al., 2017).

Alwin (2011) criticized the MTMM approach based on three considerations. First, the MTMM approach as implemented by Saris and his colleagues usually presents the question formats (methods) in a single order: it therefore assumes this ordering does not impact the results. Second, the MTMM approach does model correlated error, but assumes that all such errors are caused by systematic response styles related to the question formats varied. The plausibility of this assumption strongly depends on which experiments were performed. Third, the MTMM approach assumes respondents' random errors should be uncorrelated. This assumption may be violated when respondents do not provide an independent answer on the second measurement occasion, but simply repeat the answer they remember given on the first occasion, that is, when there are memory effects. Saris & Van Meus (1991) argued such effects would not usually occur after at least 20 min of interview time (cf. Saris, 2013), but recent preliminary work has cast doubt on this assertion (Rettig et al., 2019).

2.1 | Limitations of the one-factor-at-a-time approach

While each of the discussed broad approaches has its advantages and disadvantages, these approaches also have an—under recognized—shared drawback: either experimentally or observationally, most existing work varies 'one factor at a time' ('OFAT') to investigate the impact of response error sources on random and correlated error. OFAT is problematic for three reasons.

First, varying question forms one-at-a-time and comparing the results across studies, as in meta-analysis (e.g. Saris & Gallhofer, 2007) or literature reviews (e.g. DeCastellarnau, 2017), while clearly very useful as a first step, is statistically inefficient as an estimation strategy. The OFAT approach is well known to give lower power, larger standard errors and higher sample size requirements, compared with (fractional) factorial experimental designs (Cox, 1958; Cox & Reid, 2000).

Second, OFAT prevents us from evaluating how the importance of response errors measure up against each other. For example, when asking respondents' opinions on immigration, several studies have demonstrated the existence of acquiescence (Révilla, Saris, & Krosnick, 2014), while others have demonstrated social desirability bias (Janus, 2010). But which is more important? Where should we focus resources to improve conclusions? To answer these questions, errors should be evaluated simultaneously.

Third, when design factors interact, OFAT can mislead us regarding the best way to ask a question. In DoE terms, OFAT confounds interactions with main effects. For example, by comparing estimates of random and correlated error in immigration questions asked using an 'agree–disagree' scale, Saris et al. (2010) found, after comparing 5-point, 7-point and 11-point scales, that 5-point scales performed best, because they exhibited the least correlated error. Using agree–disagree questions, fewer scale points appeared better. But in an earlier meta-analysis of similar experiments, the opposite was found: fewer scale points gave worse performance (Saris & Gallhofer, 2007), while Saris et al., (2010) reported that the overall best question format for the immigration items was an item specific scale with 11 scale points. The authors interpreted this discrepancy as due to an interaction effect between the number of scale points and the type of question. Pooling together the available experiments performed on these questions across

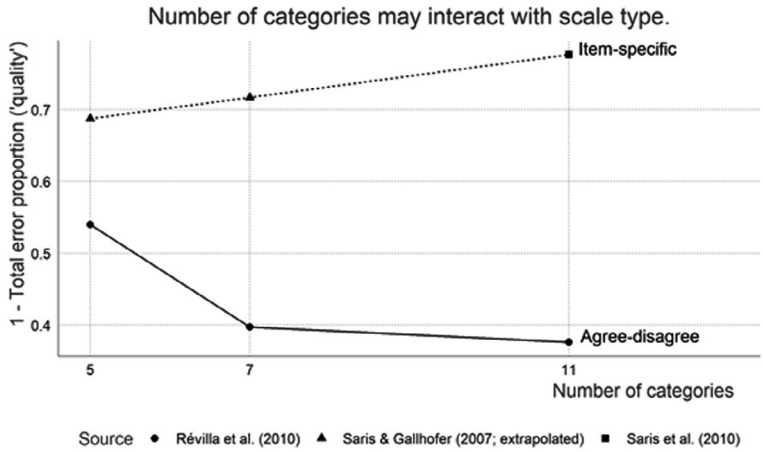

FIGURE 1 Estimated effect of number of categories and type of scale (agree-disagree vs. item specific) on proportion of 'true variance' in immigration items. Shown is the average across three immigration attitude items from multitrait-multimethod experiments performed in the ESS (circles and squares), combined with extrapolations based on a meta-analysis discussed in Saris & Gallhofer (2007; triangles). See Révilla et al. (2010) for further detail regarding these experiments

different papers gives the estimates shown in Figure 1, which do exhibit an interaction effect for these two factors. However, they were unable to test this conjecture, because the agree–disagree formulation and number of scale points were not varied simultaneously. In short, studies comparing the effect of a single design factor often do not provide the information survey researchers need to optimize their questions.

2.2 | The proposed approach: multitrait-multierror design and model

The DoE literature has emerged with the express purpose of tackling the disadvantages of the OFAT approach. Here, we propose to leverage this body of knowledge in the form of within-person fractional factorial experimental *designs*, which we then analyse using latent variable *models*. For example, if acquiescence, social desirability, method effect and random error are all thought to play a role in response error, and the researcher is able to define two-group split-ballot experiments for each of these factors, we propose instead a fractional design based on the full $2 \times 2 \times 2 \times 2$ factorial.

As in the classical split-ballot approach, MTME identifies the response error sources that are thought to be relevant to the survey questions at hand, and question forms that can manipulate these errors experimentally. As in the quasi-simplex approach, we also allow for random error, while, as in the MTMM approach, we recognize that these error sources can operate across different but similarly measured questions. Unlike existing approaches, however, we manipulate all of these error sources simultaneously by considering their experimental manipulations as factors in a within-person experiment. To reduce respondent burden, a fractional factorial design is used in which second-order interactions are unconfounded, and each respondent need only answer two forms of the same question. We randomize the order of these forms to deal with the problem of order effects. Additionally, a latent variable model is introduced that can be applied to the resulting data to inform survey methodology and question design. Our approach, thus,

allows for the estimation of mean bias, random error variance and correlated error variance from multiple error sources simultaneously—letting the researcher evaluate the relative importance of different response error sources, as well as their interactions.

To summarize, our contribution is threefold. First, we explicitly separate the *design* step and the *analysis* step in the estimation of measurement error. This enables researchers to be more theory driven in their endeavours by enabling them to develop the design needed to estimate the measurement errors of interest and to model them using LVM. More importantly, it enables them to estimate multiple types of measurement errors concurrently. In our example we show how to do that when we expect four types of measurement error: random error, social desirability, acquiescence and method effects. Such an approach is essential if we want to advance our knowledge of measurement error from a Total Error Perspective. For a more general discussion on how to develop the MTME *design* see Cernat and Oberski (2019).

Second, we explicitly introduce the theory and knowledge from the DoE literature in the estimation of measurement error. This highlights the need for more efficient designs that estimate multiple measurement errors at the same time. It also makes explicit the assumptions of designs such as MTMM and quasi-simplex which deal with measurement error OFAT. From this perspective the classical MTMM is a special case of the more general MTME that makes very stringent assumptions about the structure of measurement error.

Third, we show how we can analyse an MTME *design* using LVM. Our model is able to estimate four types of measurement errors concurrently because of the ability to combine experimental designs with latent variable modelling. We should also note that the design and the analysis presented here is just one implementation of the MTME. The real strength of the approach is the ability to map the theoretical expectations regarding measurement error to a design and a statistical model.

Note that the idea of applying latent variable models to within-person experiments has a long history. Indeed, the word 'factor' in 'factor analysis' originally referred to the same concept as the word 'factor' in 'factorial experiment'. The psychological literature on 'facet design' (Guttman, 1959), also, suggested manipulated 'facets' of a psychological scale and mapping these to a latent space, although Guttman himself preferred his own 'smallest space analysis' as a method of analysis (Guttman, 1982). In addition, the current approach shares many goals with generalizability theory (*g*-theory), for which factor-type models, including IRT models, have also been proposed (De Boeck, 2008). Just as the classic paper of Andrews (1984) contributed the idea of identifying question forms with 'methods' in the MTMM designs introduced by Campbell and Fiske (1959), our contribution should be seen as identifying response error manipulations with 'facets' introduced by Guttman or with factors in a the more modern LVM framework.

3 | SURVEY DESIGN AND DATA

In order to collect data using the MTME design we have used the UK Household Longitudinal Study – Innovation Panel (UKHLS-IP, University of Essex, Institute for Social & Economic Research, 2017). In this design we manipulated method effects, acquiescence and social desirability.

The UKHLS-IP is a national representative household longitudinal study of England, Scotland and Wales that collects data yearly (Jäckle et al., 2017). Sampling was done using the Postcode Address File and was stratified by the percentage of household heads classified as non-manual

and population density. The initial sample was clustered in 120 sectors and had 2760 addresses. Waves 4 and 7 included refreshment samples of 960 and 1560 new addresses using similar sampling procedures.

In this paper, we use data collected as part of wave 7 which was collected in May–October 2014. Data collection was carried out using either a sequential mixed mode design: Web-Computer Assisted Personal Interview (CAPI) or single mode CAPI. In wave 5, a random two thirds of respondents were allocated to the mixed mode design while a random third received the single mode. This selection was kept in wave 7. The wave 7 refreshment sample was collected using the CAPI single mode. UKHLS-IP wave 7 achieved a 78.5% conditional household response rate and an 82% conditional individual response rate. For more details regarding the study please refer to Jäckle et al., (2017).

In order to implement the MTME design we selected six questions regarding attitudes towards immigrants (Table 1). These variables have been widely used in a number of national and international studies, such as the European Social Survey. The selection of the variables was due to the sensitivity of the topic as well as the difficulty of collecting high-quality measures on attitudes (Alwin, 2007; Saris & Gallhofer, 2007).

Our experimental design manipulated three types of correlated errors:

- Social desirability (positively vs. negatively worded questions)
- Acquiescence (agree–disagree vs. disagree–agree response scale)
- Method (2-point vs. 11-point response scale)

Social desirability refers to the tendency of respondents to change their answers in order to present themselves in a more positive light (DeMaio, 1984; Tourangeau et al., 2000). Given the sensitivity of the topic, we believe this to be a potential source of error (Janus, 2010). In order to manipulate the social desirability direction of the question we have changed the body of the question to be either positively worded or negatively worded (e.g. *We should allow* **more** *people...* vs. *We should allow* **fewer** *people...*). In this way the social norm regarding attitudes towards immigrants is manipulated in the question body. If respondents wish to present themselves in a positive light they will tend to agree with the perceived social norm.

Acquiescence, or 'yea-saying', is the tendency of respondents to agree with statements regardless of the content of the questions (Billiet & McClendon, 2000; McClendon, 1991). In order to manipulate this type of tendency we have implemented either an agree–disagree response scale or a disagree–agree one. Acquiescence is a 'satisficing' process that is used to minimize cognitive effort in the response process. When the agree category is the first one available this process is facilitated, leading to higher levels of acquiescence (Billiet & McClendon, 2000).

Finally, method effect can be defined as any characteristic of the data collection that can influence the way respondents answer questions. In survey research this has been typically conceptualized as the impact of the response scale (Andrews, 1984; Saris & Gallhofer, 2007). In this design we manipulate the impact of the method effect using either a 2-point response scale or an 11-point one.

By combining the three manipulations we can conceptualize eight different **wordings** to ask the survey questions (W1 to W8 in Table 2). For example, the first wording (W1 in Table 2) uses a negative wording of the question, with a 2-point response scale ordered as agree–disagree. Wording 2 uses the same response scale but makes the question positively worded.

TABLE 2 The eight experimental wordings used in the data collection. The first question is given as example of positive and negative wording

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

AD, Agree–Disagree response scale; DA, Disagree–Agree response scale.

In order to estimate the amount of random and correlated error variance from each source, a within-person experimental design is required. A fully crossed within-person design would repeat each question eight times, once using each of the wordings shown in Table 2. This clearly is not feasible and almost certain to produce carryover effects. Instead, we implemented a reduced, fractional factorial, design in which each respondent only received two versions of the same questions, randomly. Thus, each respondent was asked the six questions regarding attitudes towards immigrants twice, once at the beginning of the survey and once at the end. The reduced design is similar to the planned missing data design used in the split-ballot MTMM approach (Saris et al., 2004). Contrary to the common

implementation of MTMM (Saris & Gallhofer, 2007), however, the order of the forms was also randomized.

4 | STATISTICAL MODELS

Our suggested approach involves both the MTME *design* and a statistical model that can estimate mean bias, random error variance and correlated error variance based on this design, the MTME *model*. Below we detail the implementation of this model.

The MTME design is a fractional factorial within-person split-ballot experiment. We propose to analyse the data from such experiments using latent variable models (LVMs; Skrondal & Rabe-Hesketh, 2004). We estimate the model by separating the reliable (trait) variance from the correlated error and random error. For each dimension manipulated in our experiment (method, acquiescence and social desirability) we estimate a latent variable that will represent that correlated error. In addition to these variances we also estimate the bias in the mean due to our experimental factors. This is done by fixing the intercepts of the observed variables at zero and freely estimating the means of the latent variables.

The loadings are restricted based on the experimental design (Tables 2 and 3). For example, all questions measuring 'allow people of the same race and ethnic group' (Table 1) will have loadings restricted to +1 (i.e. they measure this concept) in the relationship with the latent variable T_1 and 0 (i.e. they do not measure this concept) with the rest of the trait variables. Questions with the first wording (W1 in Table 2) will have their acquiescence loadings fixed to +1, because higher scale points for this wording indicate agreement, while they will have their social desirability loadings fixed to −1, because disagreement is expected to be the socially desirable response for this wording. For the method factor we use the 2-point scale as the reference so only the questions using wordings 3, 4, 7 and 8 will have loadings fixed to +1 in relationship to the method (M) factor. Thus, the method factor encodes the correlated error among 11-point questions over and above (*relative to*) that found among 2-point questions (Eid, 2000). The intercepts of the observed variables are set to zero in order to identify the means of the latent variables. This corresponds to an assumption of no third-order interactions in a regression context.

TABLE 3 Design matrix for multitrait-multierror (MTME) model measuring attitudes towards immigrants in UKHLS-IP. These are used to determine the loadings in our model (see Assumptions)

	Subscript			Trait	Method	Acquiescence	Social desirability
Wording	*m*	*a*	*s*	(T)	(M)	(A)	(S)
W1	1	1	1	1	0	1	1
W2	1	1	2	−1	0	1	−1
W3	2	1	1	1	1	1	1
W4	2	2	2	−1	1	1	−1
W5	1	2	1	1	0	−1	1
W6	1	2	2	−1	0	−1	−1
W7	2	2	1	1	1	−1	1
W8	2	2	2	−1	1	−1	−1

4.1 | Model and assumptions

Denote the observed variable on the t-th trait, m-th method, a-th acquiescence direction, and s-th social desirability direction as y_{tmas}. In our application, we observe $6 \times 2 \times 2 \times 2 = 48$ such variables. We then consider a response variable y^*_{tmas}; for binary (2-point) variables, this is the inverse probit, $y^*_{tmas} = \Phi^{-1}\left(y_{tmas} = 1\right)$ when $m = 0$, and it is the identity otherwise, $y^*_{tmas} = y_{tmas}$ when $m = 1$. Our response model for each response variable is then a linear factor model,

$$y^*_{tmas} = \lambda^{(T)^*}_{tmas} T_t + \lambda^{(M)^*}_{tmas} M + \lambda^{(A)^*}_{tmas} A + \lambda^{(S)^*}_{tmas} S + \varepsilon_{tmas}, \tag{1}$$

where the loadings $\lambda^{(T)^*}_{tmas}$, $\lambda^{(M)^*}_{tmas}$, $\lambda^{(A)^*}_{tmas}$ and $\lambda^{(S)^*}_{tmas}$ are restricted according to the corresponding element, $x_{mas,\cdot}$, of a design matrix X (Table 3),

$$\lambda^{(\cdot)^*}_{tmas} = \lambda^{(\cdot)}_m x_{mas,\cdot}, \tag{2}$$

and we allow a method-specific scaling factor $\lambda^{(\cdot)}_m$ for each latent variable, with $\lambda^{(\cdot)}_1 = 1$ for identification purposes. In our application, the scaling factor $\lambda^{(\cdot)}_m$ scales standardized effects on the probit scale (for 2-point scales) to effects on the unstandardized 11-point scale.

The design matrix used in our application is given in Table 3, while wordings corresponding to the rows of this matrix are found in Table 2. For example, for the disagree–agree question using wording W6, 'the UK should allow **more** people of the same race or ethnic group as most British people to come and live here', with 2-point scale, the corresponding subscript is $t = 1$, $m = 1$, $a = 2$, $s = 1$ (see Table 2), and after inserting the corresponding elements from the design matrix in Table 3 into Equation (2), the response model is

$$y_{1121} = T_1 - A + S + \varepsilon_{1121},$$

for comparison, the 11-point version of the same question, that is, y_{1221}, has response model

$$y_{1221} = \lambda^{(T)}_2 T_1 + M - \lambda^{(A)}_2 A + \lambda^{(S)}_2 S + \varepsilon_{1221}.$$

Note that the method factor drops out from one of the response models due to the dummy coded ('M-1') design column (zeroes in Table 3).

Assumptions

A1. Local independence (no further correlated error). Conditional on the joint latent variable vector, the observed variables are independent, $\mathbb{E}[y_{tmas} y_{t'm'a's'} | T_1, \ldots T_6, M, A, S\,[= \mathbb{E}]\, y_{tmas} \,|T_1, \ldots T_6, M, A, S\,[\mathbb{E}]\, y_{t'm'a's'} \,|\, T_1, \ldots T_6, M, A, S]$ for any of $t \neq t', m \neq m', a \neq a'$, or $s \neq s'$.

A2. Linearity (Equation 1). The indicators are linear measures of the latent traits, and the random response error variables M, A and S.

A3. Additivity of errors (no error interaction; Equation (1)). The random response error variables and latent trait are additive in their effect on the response.

A4. Homoskedasticity. The variances and covariances of all latent and observed variables are constant.

A5. Residual error independence (exogenous random errors). The random error variables have zero mean, $\mathbb{E}\left[\varepsilon_{tmas}\right] = 0$, and are uncorrelated with each other, $\mathbb{E}\left[\varepsilon_{tmas}\varepsilon_{t'm'a's'}\right] = 0$ for any differing t', m', a' or s', with the traits and with the random response error variables, $\mathbb{E}\left[\varepsilon_{tmas}r\right] = 0$, for $r \in \left\{T_{t'}, M, A, S\right\}$.

A6. Response error independence. The random response error variables are uncorrelated with each other and with the traits, $\mathbb{E}\left[rr'\right] = \mathbb{E}\left[r\right]\mathbb{E}\left[r'\right]$ for $r \in \{M, A, S\}$, $r' \in \left\{T_t, M, A, S\right\}$ and $r \neq r'$.

A7. Independent and identically distributed (i.i.d.) observations.

Of the above assumptions, A1 is the most crucial; all others can be relaxed, at least partially. Note that Assumptions A2 and A3 are already implied by the response model (1); we repeat them here for completeness. Furthermore, Assumptions A3–A4 are implied in our Bayesian model, by the multivariate normality of all random variables. We include these assumptions here because the model does not rely on distributional assumptions beyond those implied by A1–A6, even if the Bayesian approach requires such distributional assumptions. Assumption A7 may be violated in the case of complex sample surveys, and can be relaxed using methods standard in the SEM literature (e.g. Oberski, 2014).

Assumption A6 (see Saris & Gallhofer, 2007; Widaman, 1985) can greatly aid empirical identification and parameter stability (Oberski, 2019) but may be somewhat unrealistic in practice. For example, a person for whom the social desirability effect is stronger may be thought more prone to acquiescence bias as well. In our model fitting, we included assumption A6 as a baseline and investigated some relaxations of it (specifically a free correlation between S and A).

Also note that our model implicitly allows for an interaction in the effect of design choices on the quality of the indicators and as the effect of a design choice on the 'quality' (estimated correlation with T) of a question does depend on the other design choices. A3 could be released by allowing loadings to differ over more factors, although identification and estimation issues will often arise.

Part of our model is presented graphically in Figure 2. Here we only show the first survey item (T_1) measured using the eight wordings (W1–W8) from Table 2. The squares represent the observed variables, while latent variables are shown in circles. The full model includes the observed

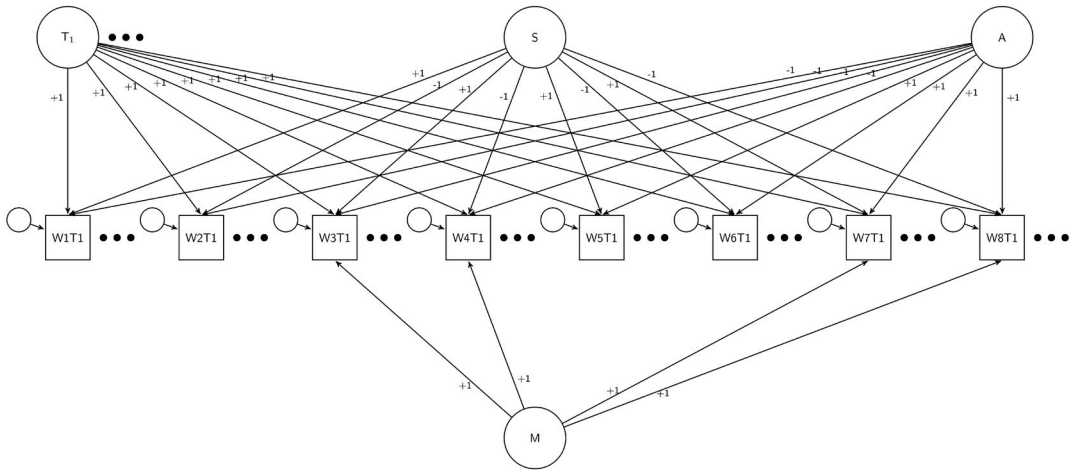

FIGURE 2 Visual representation of statistical model. Model for question 1 is shown for simplicity
63 × 27 mm (240 × 2400 DPI)

and latent variables for the other five questions as well. Thus, there are $8 \times 6 = 48$ observed variables and nine latent variables, not counting the 48 random error terms. Due to the restrictions on the loading matrix, the model has 9 mean parameters and 48 variance and covariance parameters. Although there are 48 observed variables and 2,314 cases, there is a large amount of missingness due to the fractional factorial (planned missing completely at random, MCAR) design, which provides about 70 cases for each of the 1128 pairwise correlations. We deal with missing data by formulating an ignorable likelihood, on which the Bayesian estimation is based.

4.2 | Identification

Identification of latent variable models cannot usually be established globally (for any parameter value). However, in the literature local identification, that is, the uniqueness of the maximum likelihood in an open neighbourhood, is usually deemed sufficient. For (linear) structural equation models such as ours, a sufficient condition for identification is that the Jacobian of the implied covariances with respect to the parameter vector be of full column rank (Bekker, 1989; Wald, 1950). The fractional factorial MTME design guarantees that all rows of this Jacobian are available, by guaranteeing (with probability converging to one) the observation of all first and second-order moments.

The MTME model also guarantees the fulfillment of the column rank condition 'almost surely' through the use of the (fractional) factorial design. Specifically, as shown by Bekker (1989) for confirmatory factor models, the rank of the Jacobian depends on the ranks of the loading and latent covariance matrices. The MTME restrictions outlined above guarantee that $\mathbf{\Phi}$ is full rank. The rank of the loading matrix $\mathbf{\Lambda}$ is determined by the MTME design (columns M, A and S in Table 3), which, by the definition of a fractional factorial, is full rank, and their dependence on the trait loading columns. Including the columns corresponding to latent traits, these are independent of the column S, except when the free loadings of the trait, method and social desirability factors are equal, $\lambda_2^{(T_t)} = \lambda_2^{(M)} = \lambda_2^{(S)}$. In this case, the rank deficiency $\phi_{99} = \sum_{i\in\{1,\ldots,6\}} \sum_{j\in\{i,\ldots,6\}} \phi_{ij}$ occurs; that is, the social desirability factor variance cannot be separated from the trait variances and covariances. Note that the method factor variance ϕ_{77}, acquiescence factor variance ϕ_{88} and residual error variances ψ_{ii} never suffer from this rank deficiency. Because the probability measure of the event $\lambda_2^{(T_t)} = \lambda_2^{(M)} = \lambda_2^{(S)}$ is zero, the model is said to be 'almost surely' locally identified.

In spite of 'almost sure' identification, the presence of a rank-deficient point in the parameter space can generate nonconvergence, inadmissible estimates and unstable estimates, a fact widely reported in the MTMM literature (see Oberski, 2019 and references therein). In addition, as the number of response error factors grows, so does the fraction of the MTME factorial, and thereby the fraction of missing information on the second-order moments (rows of the Jacobian). This means that in practice, increasing the number of response error factors will require the sample size to grow exponentially. Even with large samples, estimation problems may still occur due to the rank deficiency (Oberski, 2019). We have stabilized the estimates by using Bayesian weakly informative priors.

4.3 | Estimation

To estimate the model, we have used Bayesian estimation as implemented in Mplus 8 (Muthén & Muthén, 2017), and as recommended by Helm et al., (2017) for MTMM models. Bayesian estimation

was chosen for two reasons: (1) computational convenience of Gibbs sampling for the present, complex, model with relatively many latent variables and large amount of missing information; (2) the use of priors to prevent inadmissible estimates and nonconvergence commonly found in MTMM models (Helm et al., 2017), and to allow us to incorporate information from existing studies on response error variance of these survey items (Oberski et al., 2012; Saris et al., 2010).

The default distributional assumptions were used, namely $\boldsymbol{y}, \boldsymbol{\eta} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu} = \left(\boldsymbol{\Lambda}\boldsymbol{\kappa}', \boldsymbol{\kappa}'\right)'$ and covariance $\boldsymbol{\Sigma} = \begin{bmatrix} \Lambda\Phi\Lambda' + \Psi & \boldsymbol{\Lambda} \\ \Lambda' & \boldsymbol{\Phi} \end{bmatrix}$, where $\boldsymbol{y}$ collects the observed variables in a vector, $\boldsymbol{\eta}$ collects the latent variables in a vector, $\boldsymbol{\Lambda}$ is the loading matrix, $\boldsymbol{\kappa}$ is the latent mean vector, $\boldsymbol{\Phi}$ is the latent variance matrix and Ψ is the residual variance matrix.

The priors used for the loading parameters were uninformative, $\lambda_2^{(\cdot)} \sim N(0, 100)$. For the variance parameters of the response errors, we used somewhat informative priors,

$$\text{Var}(M) = \varphi_{77} \sim \text{IG}(2, 1) \quad \text{Var}(A) = \varphi_{88} \sim \text{IG}(2, 1), \quad \text{Var}(S) = \varphi_{99} \sim \text{IG}(2, 1), \tag{3}$$

which centre the expected variance contribution of these factors on unity and have infinite variance. These priors were chosen based on results from the literature regarding response error variance in these survey items. For the variance–covariance matrix of the latent traits, we used $\boldsymbol{\Phi}_{1:6,1:6} \sim \text{IW}(0, 10)$, where the inverse Weibull is slightly informative but very close to the default 'uninformative' prior, IW (0, 7) (Asparouhov & Muthén, 2010, 2010b). For the residual variances and latent means, the default uninformative priors, inverse Gaussian and normal, were used, $\Psi_{ii} \sim \text{IG}(-1, 0)$ and $\boldsymbol{\kappa}_i \sim \text{N}(0, \infty)$. We used the default PX1 Gibbs sampler (Asparouhov & Muthén, 2010b), with four chains and 200,000 iterations.

5 | RESULTS

The model converged and the overall credible interval for the chi-square statistic was between 56.4 and 350.1 with 63 free parameters and a posterior predictive $p = 0.004$. The trace plots and posterior distributions did not indicate convergence issues (trace plots and posterior distributions are provided as an online appendix). The traceplots show a good mix and the four chains consistently overlap. The potential scale reduction factor (Gelman & Rubin, 1992) was 1.019 and we have eliminated the first 100,000 iterations (burn-in). The final model is also consistent with results from a model run for 100,000 iterations.

The MTME analysis can be summarized in different ways. One way to understand the results of the analysis is to investigate the variances and the means of the latent variables estimated (Table 4). Of special importance in our model are the mean and variance estimates for the three types of correlated errors that we have manipulated: Acquiescence, social desirability and method. If measurement error is absent, we expect the means and variances of these latent variables to be close to zero.

First, we investigate the impact of the factors manipulated in our experiment on the means of the observed variables (Table 4). The choice of agree–disagree versus disagree–agree (acquiescence) changes the observed mean of the variables by 0.25 standard deviations, regardless of the question wording or the response scale. Similarly, using either positively or negatively worded questions (social desirability) impacts the means of the observed variables by 0.18 standard deviations. Last, the expected mean for questions that use 11-point scales is 5.13 higher compared to

TABLE 4 Mean and variance estimates for the latent variables of the multitrait-multierror (MTME)

	Mean			Variances		
Latent variable	**Point estimates**	**Lower CI**	**Upper CI**	**Point estimates**	**Lower CI**	**Upper CI**
Traits						
Allow same race	−0.42	−0.60	−0.17	4.09	3.25	5.03
Allow different race	−0.98	−1.18	−0.72	5.96	4.91	7.12
Allow poorer countries	−0.97	−1.17	−0.71	5.53	4.53	6.67
Good for economy	0.24	0.04	0.50	8.56	7.14	10.19
Culture enriched	0.50	0.29	0.77	9.44	7.86	11.20
Better place to live	0.02	−0.19	0.28	9.25	7.73	10.95
Correlated errors						
Acquiescence	0.25	0.19	0.31	0.42	0.30	0.56
Social desirability	−0.18	−0.40	−0.09	0.30	0.14	0.69
Method (11 pt)	5.13	5.04	5.22	0.87	0.68	1.11

CI, credible interval.

the 2-point scales. If method effect would have no impact on the observed mean we would have expected a value of 5.5 ($11/2 = 5.5$). In our results the mean is significantly lower, implying that people tend to underestimate their attitudes towards immigrants by 0.27 when using an 11-point response scale compared to a 2-point scale.

These results can be concerning for survey methodologists as relatively small wording changes can lead to important shifts in the means. That being said, we cannot say which wording has less mean bias without making some assumptions such as the 'more is better' one used in survey methodology to investigate social desirability.

In addition to the mean biases introduced by measurement error, correlated error variance also plays an important role. We can see that all three latent variables have substantial levels of correlated error variance (and credible intervals do not include 0). The highest variance is due to the method, followed by acquiescence and social desirability.

Another way to understand the impact of the correlated errors on the variance of our observed questions is to decompose the total variance. This is done by calculating the percentage of variation explained by each component out of the total variance. Figure 3 presents the six questions as well as the eight wordings for each (see Table 2). Within each one we can see the total valid variance (trait), as well as the other sources of measurement error. Again, typical substantive analyses might assume perfect measurement, that is, trait represents 100% of the observed variation. Here, this assumption is always wrong although the degree to which this is the case varies by question and wording.

Figure 3 highlights a number of patterns. First, wordings 1, 2, 5 and 6 have the most amount of valid variance. For all of these wordings a 2-point response scale was used. It is thus obvious that the 11-point scale produces more correlated errors as compared to the 2-point scale. Furthermore, we observe that random error is the main source of invalid variance regardless of the response scale used. Nevertheless, we also observe important amounts of variation due to acquiescence, social desirability and method, especially when the 11-point scale is used. Lastly, it appears that the first question, and to a lesser degree the second and third ones, have more random error compared to the last three questions.

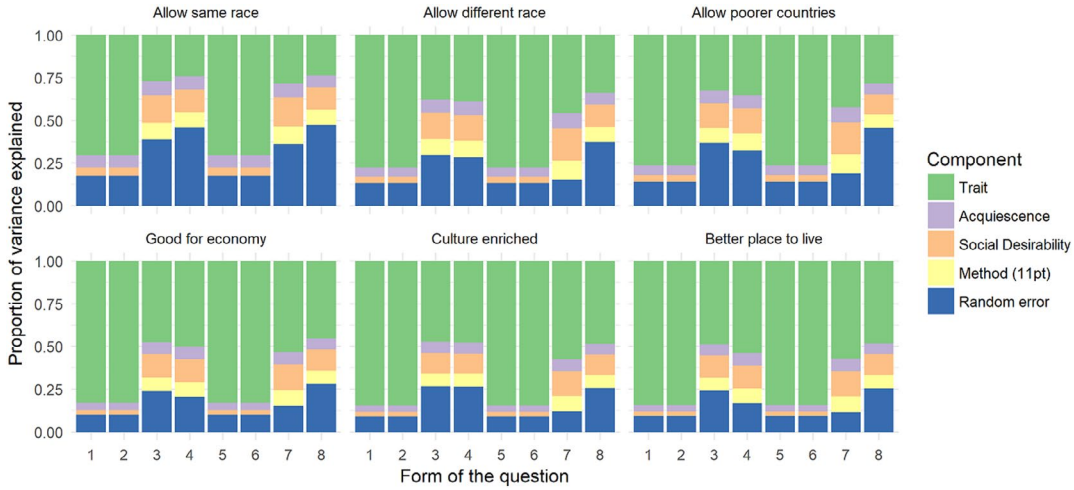

FIGURE 3 Variance decomposition of observed variables by question and wording [Colour figure can be viewed at wileyonlinelibrary.com]

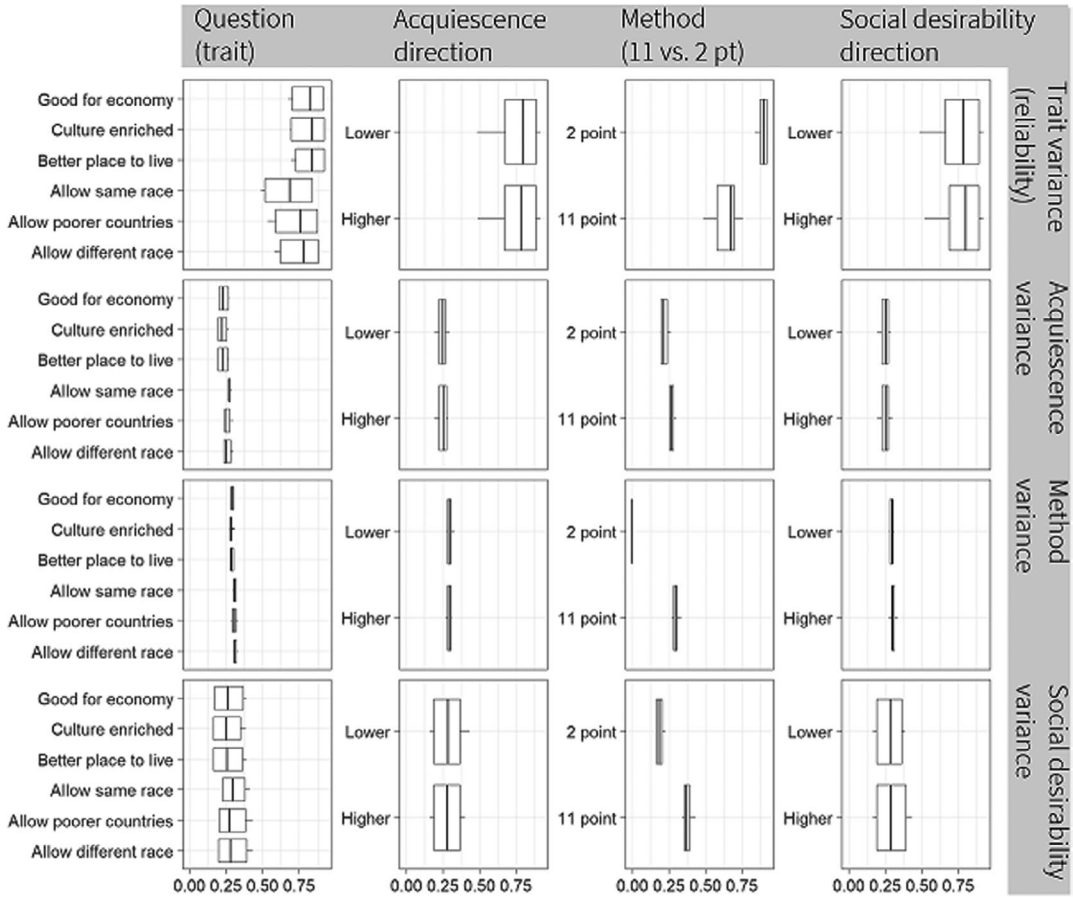

FIGURE 4 Meta-analysis based on multitrait-multierror which shows the relationship between the observed and the latent variables

Yet another way to understand the results from the MTME is by averaging the relationships between the observed variables and the latent ones. Figure 4 presents the aggregated results where the horizontal axis represents the strength of the relationship between the observed and the latent variables (based on standardized loadings). The columns represent dimensions we have manipulated (questions, acquiesce, method and social desirability), while the rows represent the latent variables estimated. For example, the first row in Figure 4 shows the degree to which reliability varies by the question, by the acquiescence direction, by the number of response categories and by social desirability direction. We can observe that reliability (relationship between the observed variables and the trait variables) tends to be similar between the different types of questions but on average questions with 2-point response scales are more reliable than those with 11-point response scales (in line with Figure 2). Figure 4 highlights that not only are questions with 11-point scales more unreliable but they also have more method and social desirability bias.

6 | DISCUSSION AND CONCLUSIONS

In this paper, we have proposed a new way to design and estimate multiple types of measurement error concurrently: the MTME approach. Using the UKHLS-IP we have implemented an MTME design that estimates social desirability, acquiescence, method effect, as well as random error.

We have shown that these types of measurement errors impact both the means and the variances of the observed variables. Furthermore, we have seen that they can be relatively large, leading to validity coefficients close to 0.3 for some of the questions. One of the main causes for this variation in data quality was the method effect, or response scale used. Our results highlight that the 2-point response scale shows important differences in their reliability, validity and means compared to the 11-point scale.

We also acknowledge some of the limitations of the proposed method. One of the most important ones is the possible memory effect that can be present between the two applications of the question forms/wordings. This limitation, which is present in all within experimental design, has already attracted considerable attention in the survey research literature (Alwin, 2011; Krosnick, 2011; Saris & Gallhofer, 2007). In this design we have aimed to ameliorate this by randomizing the form order as well as imposing a minimum period between the two forms of 5 min (average time between forms was 30 min). Further research is needed to see how the results of the MTME might be biased because of memory effects. Another limitation of our study refers to the degree to which our manipulation of social desirability has been effective. Future research can investigate different approaches to experimentally manipulate social desirability using survey questions.

There are three main reasons why we want to better understand measurement error in survey research. First, if we can estimate multiple types of measurement error concurrently, we can truly understand the relative importance of each and can develop better ways to collect data. Second, if we can estimate multiple types of measurement error, we can also correct for them. Third, we can better understand the relationship between measurement error and other sources of bias, such as non-response. While in this paper we have mainly concentrated on the first issue, estimating multiple types of measurement error, the MTME can also be used to tackle the other two issues. For example, instead of using the observed answers to estimate the relationship between attitudes towards immigration and variables of interest, such as political affiliation, we could use the latent trait variables. Similarly, the latent variables estimating acquiescence or

social desirability could be correlated with non-response in future waves to understand if there are relationships between measurement error and selection bias.

As such, we believe that MTME represents an important advance in survey research. It tackles one of the main limitations of previous research regarding measurement error: the separate analysis of different measurement errors. By concurrently estimating multiple types of correlated error and their impact on means and variances we can get one step closer to a method of estimating total survey error, including different errors sources simultaneously.

Furthermore, the MTME can be considered a very general approach to estimation and correction of measurement error. As such, different types of designs can be implemented in computer assisted data collection methods depending on the questions and the types of measurement error that are expected. For example, if researchers believe that only one type of measurement error is present, such as social desirability, then a MTME design could be implemented to manipulate only that dimension. Similarly, other types of measurement errors could be included independently or concurrently, such extreme response style, middle response style or recency effects. Finally, the model could be estimated in different ways. For example it could be reformulated as a true score model (Eid, 2000) or as multilevel model to help with estimation or interpretation.

As such, researchers should actively consider the main types of measurement error expected in their key questions and develop appropriate experiments to estimate them. MTME designs together with latent variable modelling can provide a general framework for implementing and analysing them.

ORCID

Alexandru Cernat https://orcid.org/0000-0003-2176-1215
Daniel L. Oberski https://orcid.org/0000-0001-7467-2297

REFERENCES

Alwin, D.F. (2007) *Margins of error*. Hoboken, NJ, USA: John Wiley & Sons Inc.

Alwin, D.F. (2011) Evaluating the reliability and validity of survey interview data using the MTMM approach. In: Madans, J., Miller, K., Maitland, A. & Willis, G. (Eds.) *Question evaluation methods: contributing to the science of data quality*, 1st edition. Wiley, pp. 265–294.

Alwin, D.F. & Krosnick, J.A. (1991) The reliability of survey attitude measurement: the influence of question and respondent attributes. *Sociological Methods & Research*, 20(1), 139–181. https://doi.org/10.1177/0049124191020001005

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

Asparouhov, T. & Muthén, B. (2010) Bayesian analysis of latent variable models using Mplus. Available from Mplus website, http://www.statmodel.com/download/BayesAdvantages18.pdf

Asparouhov, T. & Muthén, B. (2010b) Bayesian analysis using Mplus: technical implementation. Available from Mplus website, https://www.statmodel.com/download/Bayes3.pdf

Bekker, P.A. (1989) Identification in restricted factor models and the evaluation of rank conditions. *Journal of Econometrics*, 41(1), 5–16.

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

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73(4), 533–559. https://doi.org/10.1007/s11336-008-9092-x

Boeschoten, L., Oberski, D. & de Waal, T. (2017) Estimating classification errors under edit restrictions in composite survey-register data using multiple imputation latent class modelling (MILC). *Journal of Official Statistics*, 33(4), 921–962.

Bollen, K. (1989) *Structural equations with latent variables*. Hoboken: Wiley-Interscience Publication.

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

Cernat, A. & Oberski, D.L. (2019). Extending the within-persons experimental design: the multitrait-multierror MTME approach. In: Lavrakas, P., Traugott, M., Kennedy, C., Holbrook, A., de Leeuw, E. & West, B. (Eds.) *Experimental methods in survey research*. New Jersey: Wiley, pp. 481-500.

Cernat, A. & Sakshaug, W.J. (in press). *Measurement error in longitudinal data*. Oxford: Oxford University Press.

Cox, D.R. (1958) *Planning of experiments*. New York: Wiley.

Cox, D.R. & Reid, N. (2000) *The theory of the design of experiments*. Boca Raton, FL: Chapman & Hall/CRC.

Czitrom, V. (1999) One-factor-at-a-time versus designed experiments. *The American Statistician*, 53(2), 126–131. https://doi.org/10.1080/00031305.1999.10474445

DeCastellarnau, A. (2017) A classification of response scale characteristics that affect data quality: a literature review. *Quality & Quantity*, 1–37.

DeMaio, T. (1984) Social desirability and survey measurement: a review. In: Turner, C. & Martin, E. (Eds.) *Surveying subjective phenomena*. New York: Russell Sage Foundation, pp. 257–282.

Eckman, S., Kreuter, F., Kirchner, A., Jäckle, A., Tourangeau, R. & Presser, S. (2014) Assessing the mechanisms of misreporting to filter questions in surveys. *Public Opinion Quarterly*, 78(3), 721–733. https://doi.org/10.1093/poq/nfu030.

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

Fisher, R.J. & Katz, J.E. (2000) Social-desirability bias and the validity of self-reported values. *Psychology and Marketing*, 17(2), 105–120.

Fuller, W.A. (1987) *Measurement error models*. Hoboken, NJ: Wiley-Interscience.

Gelman, A. & Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472. https://doi.org/10.1214/ss/1177011136

Groves, R.M. & Lyberg, L. (2010) Total survey error: past, present, and future. *Public Opinion Quarterly*, 74(5), 849–879.

Guttman, L. (1959) Introduction to facet design and analysis. *Acta Psychologica*, 15, 130–138.

Guttman, L. (1982) Facet theory, smallest space analysis, and factor analysis. *Perceptual and Motor Skills*, 54(2), 491–493.

Hagenaars, J.A. (2018) Confounding true and random changes in categorical panel data. In: Giesselmann, M., Golsch, K., Lohmann, H. & Schmidt-Catran, A. (Eds.) *Lebensbedingungen in deutschland in der längsschnittperspektive*. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 245–266.

Heise, D.R. (1969) Separating reliability and stability in test-retest correlation. *American Sociological Review*, 34(1), 93–101.

Helm, J.L., Castro-Schilo, L. & Oravecz, Z. (2017) Bayesian versus maximum likelihood estimation of multitrait-multimethod confirmatory factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(1), 17–30.

Jäckle, A., Gaia, A., Al Baghal, T., Burton, J. & Lynn, P. (2017) *Understanding society – The UK household longitudinal study, innovation panel, waves 1–9, user manual*. Colchester: University of Essex.

Janus, A.L. (2010) The influence of social desirability pressures on expressed immigration attitudes*: social desirability pressures and immigration attitudes. *Social Science Quarterly*, 91(4), 928–946.

Jöreskog, K.G. (1978) Structural analysis of covariance and correlation matrices. *Psychometrika*, 43(4), 443–477. https://doi.org/10.1007/BF02293808

Klein, R.A., Ratliff, K.A., Vianello, M., Adams, R.B., Bahník, Š., Bernstein, M.J. et al. (2014) Investigating variation in replicability: a 'Many Labs' replication project. *Social Psychology*, 45(3), 142–152.

Kreuter, F., Muller, G. & Trappmann, M. (2010) Nonresponse and measurement error in employment research: making use of administrative data. *Public Opinion Quarterly*, 74(5), 880–906. https://doi.org/10.1093/poq/nfq060

Kreuter, F., Presser, S. & Tourangeau, R. (2008) Social desirability bias in CATI, IVR, and web surveys: the effects of mode and question sensitivity. *Public Opinion Quarterly*, 72(5), 847–865. https://doi.org/10.1093/poq/nfn063

Krosnick, J.A. (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3), 213–236.

Krosnick, J.A. (2011) Experiments for evaluating survey questions. In: Madans, J., Miller, K., Maitland, A. & Willis, G. (Eds.) *Question evaluation methods: contributing to the science of data quality*, 1st edition. Wiley, pp. 265–294.

Krosnick, J.A. & Presser, S. (2010) Question and Questionnaire Design. In: Marsden, P.V. & Wright, J.D. (Eds.) *Handbook of survey research (Second edition)*. Emerald.: Bingley.

Mayer-Schönberger, V. & Cukier, K. (2014) *Big data: a revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt: Mariner Books.

McClendon, M. (1991) Acquiescence and recency response-order effects in interview surveys. *Sociological Methods & Research*, 20(1), 60–103.

Muthén, L.K. & Muthén, B.O. (2017) *Mplus user's guide*, 8th edn. Los Angeles, CA: Muthén & Muthén, p. F02294210.

Nunnally, J.C. & Bernstein, I.H. (1994) *Psychometric theory*, 3rd edn. New York: McGraw-Hill.

Oberski, D. (2014) lavaan.survey: an R package for complex survey analysis of structural equation models. *Journal of Statistical Software*, 57(1).

Oberski, D.L. (2015) Questionnaire science. In: *The Oxford handbook of polling and survey methods*. https://doi.org/10.1093/oxfordhb/9780190213299.013.21

Oberski, D.L.. (2019). Rank-deficiencies in a reduced information latent variable model. ArXiv:1911.00770 [Math, Stat]. Available from: http://arxiv.org/abs/1911.00770

Oberski, D.L., Kirchner, A., Eckman, S. & Kreuter, F. (2017) Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association*, 112(520), 1477–1489.

Oberski, D.L., Weber, W. & Révilla, M. (2012). The effect of individual characteristics on reports of socially desirable attitudes toward immigration. In Salzborn, S., Davidov, E. & Reinecke, J. (Eds.), *Methods, theories, and empirical applications in the social sciences*. pp. 151–157. https://doi.org/10.1007/978-3-531-18898-0_19

Pankowska, P.K.P., Oberski, D.L. & Pavlopoulos, D. (2018) The effect of survey measurement error on clustering algorithms, Presented at the Big Data meets Survey Science 2018. Available from: https://research.vu.nl/en/publications/the-effect-of-survey-measurement-error-on-clustering-algorithms

Rettig, T., Karem, J. & Blom, A. (2019) *Recalling survey answers: a comparison across question types and different levels of online panel experience*. Utrecht: Presented at the Utrecht.

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

Roots, A., Masso, A. & Ainsaar, M. (2016) Measuring Attitudes towards Immigrants: validation of Immigration Attitude Index Across Countries. Presented at the Understanding key challenges for European societies in the 21st century, Lausanne, Switzerland.

Saris, W.E. (1988) *Variation in response functions: a source of measurement error in attitude research*. Amsterdam: Sociometric Research Foundation.

Saris, W. (2013) Is there anything wrong with the MTMM approach to question evaluation? *Pan-Pacific Management Review*, 16(1), 47–77.

Saris, W.E. & Gallhofer, I.N. (2007) *Design, evaluation, and analysis of questionnaires for survey research*, 1st edition. Hoboken: John Wiley & Sons.

Saris, W.E., Oberski, D.L., Revilla, M., Zavalla, D., Lilleoja, L., Gallhofer, I. & et al. (2011) The development of the program SQP 2.0 for the prediction of the quality of survey questions. RECSM Working, 23.

Saris, W., Revilla, M., Krosnick, J.A. & Shaeffer, E.M. (2010) Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, 4(1), 61–79. https://doi.org/10.18148/srm/2010.v4i1.2682

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

Saris, W. & Van Meurs, A. (1991) *Evaluation of measurement instruments by meta-analysis of multitrait multimethod studies*. North-Holand.

Schaeffer, N.C. & Presser, S. (2003) The science of asking questions. *Annual Review of Sociology*, 29(1), 65–88.

Schuman, H. & Presser, S. (1981) *Questions and answers in attitude surveys: experiments on question form, wording, and context*. New York: Academic Press.

Schwarz, N., Hippler, H.-J., Deutsch, B. & Strack, F. (1985) Response scales: effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, 49(3), 388–395.

Sinibaldi, J., Durrant, G.B. & Kreuter, F. (2013) Evaluating the measurement error of interviewer observed paradata. *Public Opinion Quarterly*, 77(S1), 173–193.

Skrondal, A. & Rabe-Hesketh, S. (2004) *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. Boca Raton: Chapman & Hall/CRC.

Smith, D.H. (1967) Correcting for social desirability response sets in opinion-attitude survey research. *Public Opinion Quarterly*, 31(1), 87.

Spector, P.E., Rosen, C.C., Richardson, H.A., Williams, L.J. & Johnson, R.E. (2019) A new perspective on method variance: a measure-centric approach. *Journal of Management*, 45(3), 855–880. https://doi.org/10.1177/0149206316687295

Tourangeau, R., Rips, L.J. & Rasinski, K. (2000) *The psychology of survey response*. Cambridge: Cambridge University Press.

University of Essex, Institute for Social and Economic Research. (2017) *Understanding society: innovation panel, waves 1–9, 2008–2016*. Colchester, Essex: UK Data Archive.

Wald, A. (1950) Note on the identification of economic relations. *Statistical Inference in Dynamic Economic Models*, 10, 238–244.

Widaman, K.F. (1985) Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, 9(1), 1–26.

Wiley, J. & Wiley, M. (1974) A note on correlated errors in repeated measurements. *Sociological Methods & Research*, 3(2), 172–188. https://doi.org/10.1177/004912417400300202

How to cite this article: Cernat, A. & Oberski, D.L. (2022) Estimating stochastic survey response errors using the multitrait-multierror model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185, 134–155. https://doi.org/10.1111/rssa.12733