National Centre for Research Methods Working Paper

2/18

Estimating stochastic survey response errors using the multitrait-multierror model

Alexandru Cernat (University of Manchester)

Daniel Oberski (Utrecht University)

NCRM
National Centre for Research Methods

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

Estimating stochastic survey response errors using the multitrait-multierror model

Alexandru Cernat (University of Manchester)

Daniel Oberski (Utrecht University)

Abstract

Response errors of different types, including acquiescence, social desirability, and random error, are well-known to be present in surveys simultaneously and to bias substantive results. Nevertheless, most methods developed to estimate and correct for such errors concentrate on a single error type at a time. Consequently, estimation of response errors is inefficient and their relative importance unknown. Furthermore, if multiple potential errors are not evaluated simultaneously, questionnaire pretests may give the wrong answer regarding the best question form. In this paper, we propose a new method to estimate and control for multiple types of errors concurrently, which we call the "multitrait-multierror" (MTME) approach. MTME combines the theory of experimental design with latent variable modeling to efficiently estimate response errors of different types simultaneously and evaluate which are most impactful on a given question. We demonstrate the usefulness of our method using six commonly asked questions on attitudes towards immigrants in a representative UK study. For these questions, method effect (11-point vs. 2-point scales) was one of the largest response errors, impacting both reliability as well as the size of social desirability.

Keywords: *measurement error, experimental design, latent variable modeling, multitrait-multierror, attitudes towards immigrants*

Introduction

Survey questions remain the primary instrument that pollsters use to tap into public opinion, governments to count their citizens, and social scientists to study thoughts, feelings, and behavior. Questions, however, are subject to *response errors* (Alwin & Krosnick, 1991; Alwin, 2007; Saris & Gallhofer, 2007): systematic and random deviations of recorded answers from the truth. These errors may take a variety of well-known forms, including acquiescence (Krosnick & Presser, 2010; Eckman et al., 2014), social desirability (Smith, 1967; Fisher & Katz, 2000; Kreuter, Presser, & Tourangeau, 2008), recency or primacy (Krosnick & Presser, 2010), and extreme response (Greenleaf, 1992; Liu, Lee, & Conrad, 2015). In addition, response errors occur when any stage of the cognitive response process (Tourangeau et al., 2000) is "satisficed" rather than "optimized" by the respondent (Krosnick, 1991). For example, when asked the typical amount they have spent on buying rings during their life, married respondents might engage an availability heuristic (Tversky & Kahneman, 1973) and report only on their wedding ring – thereby biasing the estimate upward on average.

As remarked by Groves & Lyberg (2010), response errors may bias both the means and the (co)variances of the answers. First and most clearly, if married respondents overestimate the amount spent, the observed mean will differ from the true mean, causing "systematic error". Second, and perhaps more subtly, if only married respondents overestimate, there will be a variation among respondents that is unrelated to true spending, or "random error". Since random error cannot be explained by substantive factors that explain true spending, they can attenuate correlations of interest (Spearman, 1904), bias regression coefficients – down- or upwards (Fuller, 1987) - and cause the false appearance of change over time (Hagenaars, 2018). Third and finally, if the researcher correlates "amount spent on rings" to, say, "amount spent on dresses", married respondents will make similar errors on both, causing the two variables to share common error variance and correlate spuriously. Such "correlated error" [1] (Andrews, 1984) biases covariances and can wreak havoc with multivariate statistical analyses (e.g. Spector et al., 2017). In short, response errors of various types potentially cause bias in a wide range of survey analyses of substantive interest.

Investigations of response bias have a rich history. Systematic errors have received the most attention (e.g. Schuman & Presser, 1981); overviews are found in Schaeffer & Presser (2003), Krosnick & Presser (2010), and Oberski (2015). Generally, this work estimates systematic error by "split-ballot": it compares the averages yielded by different randomly administered question forms among each other, or to a "gold standard" record (e.g. Kreuter et al., 2010; Sinibaldi, Durrant, & Kreuter, 2013). Typically, a split-ballot study investigates a single response error. Random and correlated measurement error has generated a more modest body of evidence (for a recent review, see DeCastellarnau, 2017). Alwin (2007) applies latent variable longitudinal ("quasi-simplex") models to US panel data, and evaluates the extent of random error estimated across question forms. However, the random error source is left unspecified, while correlated error is assumed non-existent - an assumption criticized by Saris (2012). Saris & Gallhofer, (2007) and Saris et al. (2011) applied "multitrait-multimethod" (MTMM) models – a type of crossed random effect or hierarchical model – to within-persons experiments on the question form; they also provided meta-analyses of the results across many countries. The MTMM approach does incorporate both random and

[1] The term "correlated error", introduced by Andrews (1984), is not to be confused with "correlated response variance", i.e. correlations among *observations* rather than variables, for instance due to being interviewed by the same person (Groves, 2004, p. 26). Other terms with similar meanings sometimes employed are "common method variance" (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), "method effect", "method variance", "invalidity", and even "systematic error" (Saris & Gallhofer, 2007). To avoid confusion, we employ the term "correlated error" here, taking it to refer to a source of correlations among errors.

correlated errors but assumes zero carryover and order effects – both assumptions that were criticized by Alwin (2011). Again, MTMM leaves the sources of the random and correlated errors unspecified. Implicitly, then, both the quasi-simplex and the MTMM approaches effectively aggregate different types of response errors into either "random" or "correlated" error; this precludes us from separating out the offending types of response error.

In short, while useful, existing approaches do not distinguish different types of response errors, but focus on a single response error at a time. This is problematic for three reasons. First, varying question forms one-at-a-time and comparing the results across studies, as in meta-analysis or literature reviews, is extremely statistically inefficient. As classical results from the theory of design of experiments (DoE) show, the one-at-a-time approach leads to lower power and higher standard errors, or much higher sample size requirements, compared with (fractional) factorial experimental designs (Cox, 1958; Cox & Reid, 2000). Second, the one-at-a-time approach makes it impossible to evaluate the *relative* importance of response errors. For example, when asking opinions about immigration, several studies have demonstrated the existence of acquiescence (Revilla, Saris, & Krosnick, 2014), while others have demonstrated social desirability bias (Janus, 2010). But what is the relative extent of these problems? Which is more important? To answer these questions, a method is needed that studies these errors simultaneously. Third, the one-at-a-time approach can mislead us regarding the best way to ask a question. This is because it effectively assumes zero interactions among the response errors; when such interactions are, in fact, present, an inferior question form can appear better simply because it has been examined under suboptimal circumstances. For example, extreme response style cannot occur with only two response options, meaning that offering two response options will spuriously appear better if extreme response is stimulated by a substandard question formulation. In DoE terms, interactions are confounded with main effects in existing approaches.

In this paper, we propose to extend existing approaches by combining fractional factorial within-persons designs with latent variable models. As our models are most similar to those of the multitrait-multimethod (MTMM) approach, but evaluate several errors simultaneously – while relaxing some of its other assumptions – we refer to our approach as "multitrait-multi*error*" (MTME). 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-persons 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 leverage the resulting data to aid survey methodology and question design. Our approach, thus, allows for the estimation of systematic, random, and correlated error from multiple error sources simultaneously – thereby increasing the efficiency of the survey-methodological literature, and letting the researcher evaluate the relative importance of different response error sources, as well as their interactions.

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 and Economic Research, 2017). This is a national representative household longitudinal study of England, Scotland and Wales that collects data yearly (Jäckle, Gaia, Al Baghal, Burton, & Lynn, 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 the user guide (Jäckle, Gaia, Al Baghal, Burton, & Lynn, 2017).

In order to implement the MTME design we have 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).

Table 1. The six questions used to measure attitudes towards immigration

Trait number	Item formulation
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

There are three types of correlated errors that were manipulated in our experimental design:

- 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 of changing their answers in order to present themselves in a more positive light (DeMaio, 1984; Tourangeau, Rips, & Rasinski, 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 (see Table 2).

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.

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 ways to ask the survey questions (Table 2). For example, one wording (W1 in Table 2) could use a negative wording of the question, with a 2 point response scale ordered as agree-disagree. Wording 2 would use the same response scale but make 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 UK 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 UK 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 UK 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 UK 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 UK 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 UK 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 UK 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 UK should allow **fewer** people of the same race or ethnic group as most British people to come and live here

In order to estimate the amount of random and correlated error variance from each source, a within-persons experimental design is required. A fully crossed within-persons 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 different versions of the same questions. 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 in spirit to planned missing data design used in the split-ballot multitrait-multimethod approach (Saris, Satorra & Coenders, 2004). Contrary to the common implementation of MTMM (Saris & Gallhofer 2007), however, the order of the forms was also randomized.

Methods

Our suggested approach involves both the MTME design – detailed in the previous section – and a statistical model that can estimate systematic, random, and correlated response errors based on this design, the MTME model. Below we detail the implementation of this model.

As described in the previous section, the MTME design is a fractional factorial within-persons split-ballot experiment. We propose to analyze the data from such experiments using latent variable models (LVMs; Skrondal & Rabe-Hesketh 2004), in which each manipulation is considered a potential source of systematic, random, and correlated error by specifying one latent variable per manipulated experimental factor. Systematic response effects are then parameterized as latent mean structure, while random and correlated errors are parameterized as latent variable variances. The means and variances of these latent variables are identified through restrictions on the loadings, which follow from the experimental design matrix. For example, an "effects-coded" factor will result in loadings that are -1 or +1, depending on the question form, while dummy-coded factors will have loadings equal to zero or one. In our implementation, we choose a combination of dummy coding (for the response scale effect) and effects coding (for the other factors), to yield a convenient interpretation. Other choices may be more convenient depending on the research questions at hand.

Note that the idea of applying latent variable models to within-persons 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) in this journal 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 in the same year or with factors in a the more modern LVM framework.

In the example application discussed here, each observed response can be decomposed into five factors: trait (T_j), Method (M_k), Acquiescence (A_l), Social Desirability (S_m) and residual (E):

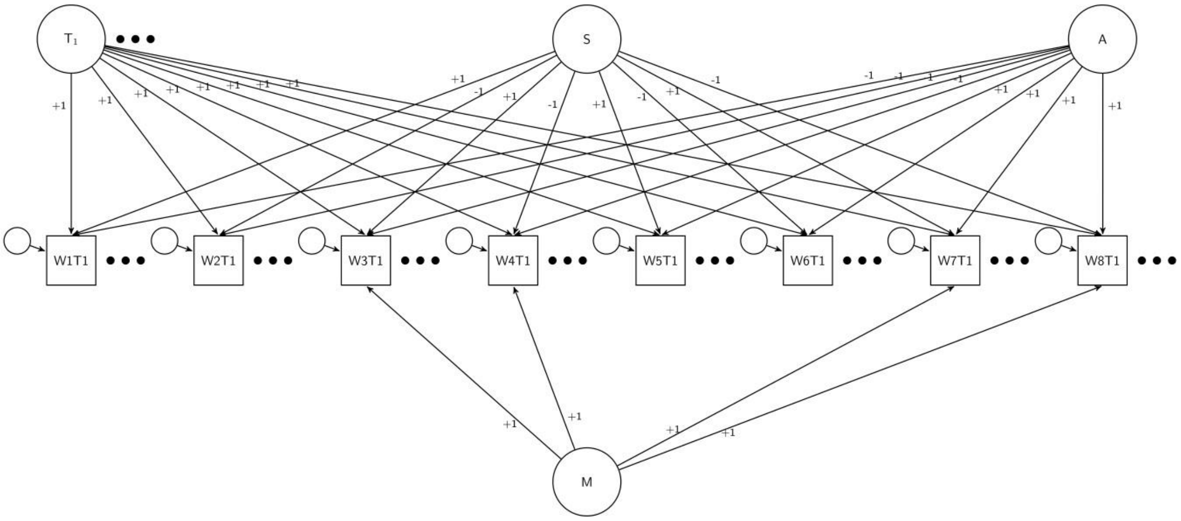
$$y_{jklm} = \lambda_{Tjklm} T_j + \lambda_{Mjklm} M_k + \lambda_{Ajklm} A_l + \lambda_{Sjklm} S_m + E_{klm}$$

The loadings (λ) are restricted based on the experimental design. For example, all questions measuring "allow people of the same race and ethnic group" (Table 1) will have loadings restricted to +1 in the relationship with the latent variable T_1 and 0 with the rest of the trait variables. All questions formatted using the first wording (Table 2): will have the loadings fixed to +1 for the relationship with the acquiescence (A) latent variable and -1 for the relationship with social desirability (S). For the method factor we use the 2 point 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. The intercepts of the observed variables are set to 0 in order to identify the means of the latent variables. This corresponds to an assumption of no third-order interactions in a regression context. The above model holds for the 11 point scales, which are analyzed as continuous response; the 2 point response scales are analyzed as categorical by formulating the above model for the probit of a positive response (Muthen, 1984).

Part of our model is presented graphically in Figure 1. Here we only show the first question (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 and latent variables for the other five questions as well. Thus, there are 8 × 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 2314 cases, there is a large amount of missingness due to the fractional factorial (planned MCAR missing data) design, which provides about 70 cases for each of the 1128 pairwise correlations. We deal with this ignorable missing data by full-information Bayesian estimation.

To estimate the model, we have used Bayesian estimation as implemented in Mplus 8 (Muthen & Muthen, 2017). We used the Gibbs sampler, with four chains and 200,000 iterations, with non-informative priors. The trace plots and posterior distributions did not indicate convergence issues (the Mplus files with trace plots and posterior distributions can be provided n request).

Figure 1. Latent variable modeling approach for the for the MTME



Results

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 3). 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 0.

First, let's investigate the impact of the factors manipulated in our experiment on the means of the observed variables (Table 3). 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. Lastly, the expected mean for questions that use 11 point scales is 5.13 higher compared to 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). As we can see, 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.

As we have argued earlier, in addition to the mean biases introduced by measurement error there is also systematic variance that is associate with it. This is essential as it can bias bivariate and multivariate analyses. We can see that all three latent variables have significant levels of correlated error (variance credibility intervals do not include 0). The highest is due to the method, followed by acquiescence and social desirability.

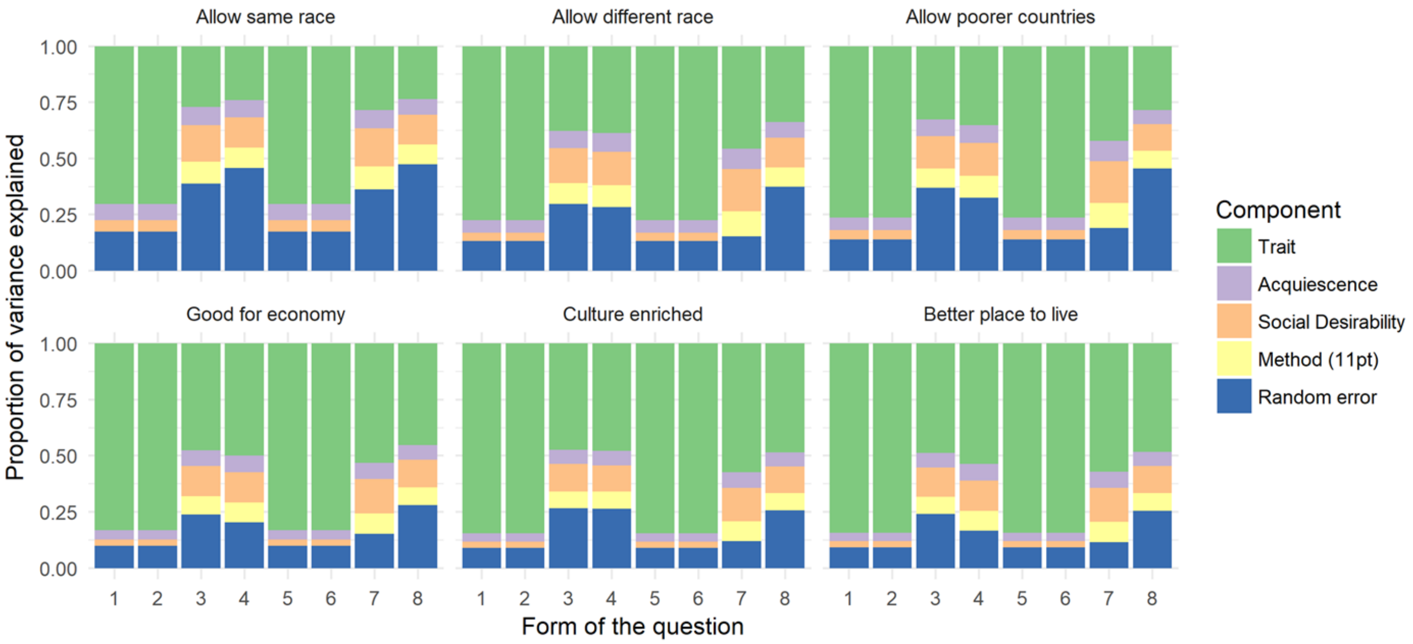
Table 3. Mean and variance estimates for the latent variables of the MTME

	Mean			**Variances**		
Latent variable	**Point estimate**	**Lower 2.5%**	**Upper 2.5%**	**Point estimate**	**Lower 2.5%**	**Upper 2.5%**
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
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

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

Figure 2 highlights a number of issues. Firstly, wordings 1, 2, 5 and 6 have the most amount of reliable variance. In all these wordings a two point response scale was used. It is thus obvious that the 11 point scale produces more correlated errors as compared to the two point scale. Furthermore, we observe that random error is the main source of unreliable variance regardless of the response scale used. Nevertheless, we also observe important amount of variation due to acquiescence, social desirability and method, especially when the 11 point scale is used.

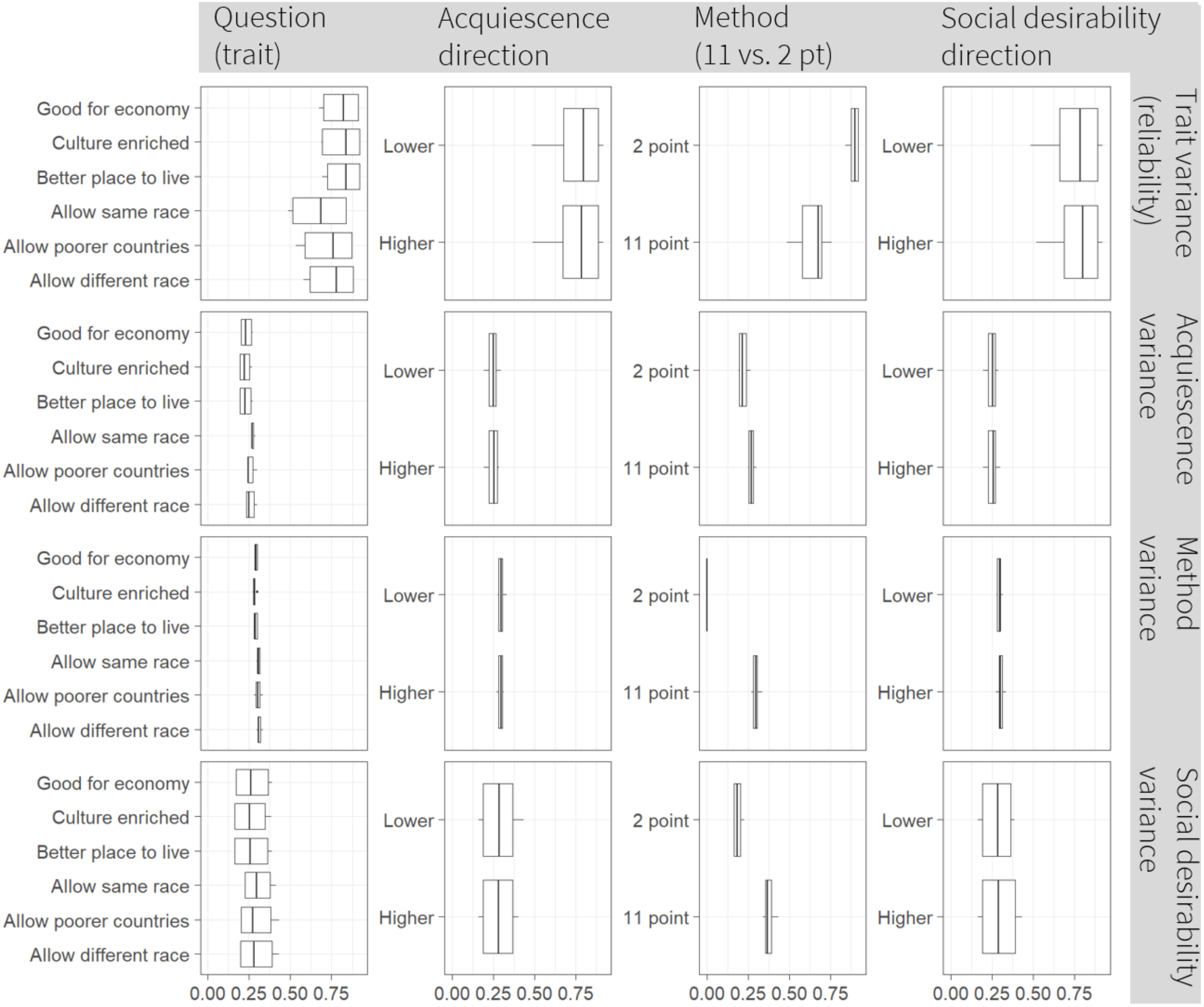
Figure 2. Variance decomposition for each question and form based on the MTME



Yet another way to understand the results from the MTME is by averaging the relationships between the observed variables and the latent ones (Figure 3). This can be conceptualized as a meta-analysis of our results. For example, the first row in Figure 3 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 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 3 highlights that not only are questions with 11 point scales more unreliable but they also have more method and social desirability bias.

Figure 3. Data quality variation by question type

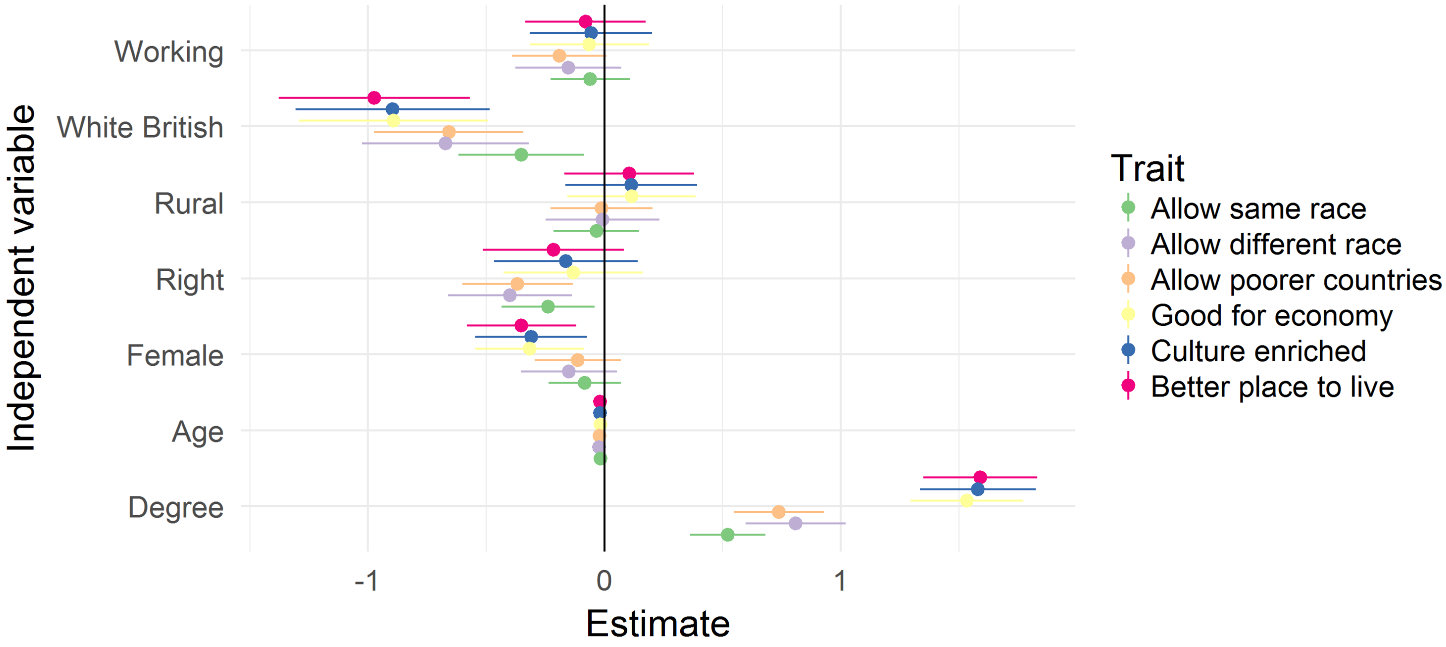


Validation

In the final step of our analysis we validate the latent variables created using the MTME model. We do this by looking at the relationships between the latent variables and criterion variables that were highlighted as important in the previous literature.

We start by looking at the six trait/questions latent variables and seven criterion variables: working, white British, rural, supports right party (conservative), female, age and degree.[2] In line with our theoretical expectation we see that white British, those that support the right and man have more negative views on immigration while having a higher degree leads to more support.

Figure 4. Validation of the trait latent variables using know criterion



Next, we investigate the relationship between the systematic measurement errors and criterion variables. The variables we use are: if the interview was face to face, an alternative acquiescence measure (based on seven agree-disagree questions), working memory (based on four questions testing working memory), if someone was present and an alternative measure of social desirability (based on scale proposed in: Steenkamp, de Jong, & Baumgartner, 2010) while we control for age, gender, and degree.

[2] For all the variables except age 1 is coded as the presence of the characteristic and 0 as the absence.

Table 4. Validation of the measurement error estimators using known criterion. Point estimated together with the lower and upper confidence interval. Bold letters are statistically significant while italics are marginally significant.

	Acquiescence			Method			Social desirability		
	Point	L 2.5%	U 2.5%	Point	L 2.5%	U 2.5%	Point	L 2.5%	U 2.5
Female	-0.01	-0.06	0.03	-0.04	-0.10	0.03	0.00	-0.03	0.03
Age	*0.00*	*0.00*	*0.00*	0.00	0.00	0.00	**0.00**	**0.00**	**0.00**
Degree	-0.02	-0.06	0.02	*0.07*	*0.00*	*0.13*	**0.07**	**0.04**	**0.10**
Memory score	-0.01	-0.04	0.03	-0.03	-0.07	0.02			
Face to face	-0.02	-0.07	0.03	-0.01	-0.09	0.07	**0.04**	**0.01**	**0.07**
Acquiescence	0.02	-0.04	0.07						
Present							-0.02	-0.06	0.02
Desirability scale							-0.03	-0.06	0.01
Right							-0.02	-0.05	0.02

We found no significant results in the relationship with criterion variables with the exception of the effect of face to face on social desirability. We also find that there is a significant and positive relationship between having a degree and social desirability. There are also marginal significant effects for age on acquiescence and degree on method effect.

Discussion and conclusions

In this paper we have proposed a new way to design, estimate and correct for multiple types of measurement error concurrently. 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 reliability 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 the number of response points used. Our results highlight that 2 point response scales have significantly higher reliability compared to the 11 point scale.

Our final step in the analysis aimed to validate the latent variables estimated using MTME. While the trait latent variables correlated as expected with part of the criterion variables the measurement error showed only partial support. This can be due to a number of reasons including the different ways in which measurement errors are estimated as well as limitations of previous research that investigated measurement error in isolation, assuming all others causes for bias are absent.

We also acknowledge some of the limitations of the 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 minutes (average time between forms was 30 minutes). 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.

Nevertheless, 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 true method of estimating Total Survey Error. 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.

References

Alwin, D. F. (2007). *Margins of Error*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9780470146316

Alwin, D. F. (2011). Evaluating the reliability and validity of survey interview data using the MTMM approach. *Question Evaluation Methods: Contributing to the Science of Data Quality*, 263–293.

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. https://doi.org/10.1086/268840

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. Boeck, P. D. (2008). Random Item IRT Models. Psychometrika, 73(4), 533–559. https://doi.org/10.1007/s11336-008-9092-x

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

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, Fla. [u.a]: Chapman & Hall/CRC.

DeCastellarnau, A. (2017). A classification of response scale characteristics that affect data quality: a literature review. *Quality & Quantity*, 1–37. https://doi.org/10.1007/s11135-017-0533-4

DeMaio, T. (1984). Social Desirability and Survey Measurement: A Review. In C. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (pp. 257–282). New York: Russell Sage Foundation.

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

Fisher, R. J., & Katz, J. E. (2000). Social-desirability bias and the validity of self-reported values. *Psychology and Marketing*, *17*(2), 105–120. https://doi.org/10.1002/(SICI)1520-6793(200002)17:2<105::AID-MAR3>3.0.CO;2-9

Fuller, W. A. (1987). *Measurement Error Models*. Hoboken, N.J: Wiley-Interscience.

Greenleaf, E. A. (1992). Measuring Extreme Response Style. *Public Opinion Quarterly*, *56*(3), 328. https://doi.org/10.1086/269326

Groves, R. M. (2004). *Survey errors and survey costs* (Repr). Hoboken, NJ: Wiley-Interscience.

Groves, R. M., & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

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 M. Giesselmann, K. Golsch, H. Lohmann, & A. Schmidt-Catran (Eds.), *Lebensbedingungen in Deutschland in der Längsschnittperspektive* (pp. 245–266). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-19206-8_14

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. https://doi.org/10.1111/j.1540-6237.2010.00742.x

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: Universi ty of Essex.

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. https://doi.org/10.1002/acp.2350050305

Krosnick, J. A., & Presser, S. (2010). Question and Questionnaire Design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (Second edition). Bingley: Emerald. Retrieved from https://pprg.stanford.edu/wp-content/uploads/2010-Handbook-of-Survey-Research.pdf

Liu, M., Lee, S., & Conrad, F. G. (2015). Comparing Extreme Response Styles between Agree-Disagree and Item-Specific Scales. *Public Opinion Quarterly*, *79*(4), 952–975. https://doi.org/10.1093/poq/nfv034

McClendon, M. (1991). Acquiescence and Recency Response-Order Effects in Interview Surveys. *Sociological Methods & Research*, *20*(1), 60–103. https://doi.org/10.1177/0049124191020001003

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49(1), 115–132. https://doi.org/10.1007/B

Muthén, L. K., & Muthén, B. O. (2017). Mplus User's Guide (Eighth Edition). Los Angeles, CA: Muthén & Muthén. F02294210

Oberski, D. L. (2015). Questionnaire Science. *The Oxford Handbook of Polling and Survey Methods*. https://doi.org/10.1093/oxfordhb/9780190213299.013.21

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

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. https://doi.org/10.1177/0049124113509605

Saris, W. E. (2012). Discussion: Evaluation Procedures for Survey Questions. *Journal of Official Statistics*, *28*(4), 537–551.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research* (Vol. 548). John Wiley & Sons.

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

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.

Schaeffer, N. C., & Presser, S. (2003). The Science of Asking Questions. *Annual Review of Sociology*, *29*(1), 65–88. https://doi.org/10.1146/annurev.soc.29.110702.110112

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

Rabe-Hesketh, S., & Skrondal, A. (2005). Multilevel and Longitudinal Modeling Using Stata (1st ed.). Stata Press.

Sinibaldi, J., Durrant, G. B., & Kreuter, F. (2013). Evaluating the Measurement Error of Interviewer Observed Paradata. *Public Opinion Quarterly*, *77*(S1), 173–193. https://doi.org/10.1093/poq/nfs062

Smith, D. H. (1967). Correcting for Social Desirability Response Sets in Opinion-Attitude Survey Research. *Public Opinion Quarterly*, *31*(1), 87. https://doi.org/10.1086/267486

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *100*(3/4), 441. https://doi.org/10.2307/1422689

Spector, P. E., Rosen, C. C., Richardson, H. A., Williams, L. J., & Johnson, R. E. (2017). A New Perspective on Method Variance: A Measure-Centric Approach. *Journal of Management*, 014920631668729. https://doi.org/10.1177/0149206316687295

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

University Of Essex. Institute For Social And Economic Research. (2017). Understanding Society: Innovation Panel, Waves 1-9, 2008-2016. Colchester, Essex: UK Data Archive. https://doi.org/10.5255/UKDA-SN-6849-9

Funding

This paper is partially financed by the National Centre for Research Methods/Economic and Social Research Council grant [R121711].

Acknowledgments

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