Survey Research Methods (2020)
Vol. 14, No. 1, pp. 79-91
doi:10.18148/srm/2020.v14i1.7450

ISSN 1864-3361
http://www.surveymethods.org

The Impact of Mixed Modes on Multiple Types of Measurement Error

Alexandru Cernat
University of Manchester

Joseph W. Sakshaug
Institute for Employment Research, and
Ludwig Maximilian University of Munich, and
University of Mannheim

Mixed mode designs are becoming standard in the collection of survey data. Despite this, there are still unknowns regarding how mode (e.g., Web) or mode design (e.g., sequential mixed mode) impacts measurement error. Previous research has been limited by the confounding of selection and measurement mode effects and the investigation of only one type of measurement error at a time. In this paper, we use three waves of the Understanding Society Innovation Panel to investigate whether single-mode versus sequential mixed-mode and Web versus face-to-face modes have different impacts on measurement error. We make use of a quasi-experimental design that randomly allocated respondents to either a unimode face-to-face interview or a sequential mixed-mode (Web and face-to-face) design. Through this design, we implement a new multitrait-multierror model that estimates social desirability, acquiescence, and method effects simultaneously. The results show no differences in measurement error between single modes and mode designs with respect to acquiescence and method effect but some differences are found for social desirability. We discuss the practical implications of these findings and their possible causes in conclusion.

Keywords: multitrait-multierror, experimental design, measurement error, mixed-mode design

1 Introduction

Using multiple modes to collect data is a common practice in survey research (De Leeuw, 2018). In particular, the practice of deploying multiple modes of data collection in sequence is widely implemented in several large-scale, policy-relevant surveys. Prominent examples include the U.S. American Community Survey, a cross-sectional survey which uses relatively inexpensive self-administered modes (mail and Web) in the initial phase of data collection, followed by more expensive interviewer-administered modes (telephone and face-to-face) during the nonresponse follow-up phase (U.S. Census Bureau, 2014). Another high-profile example is Understanding Society—the UK Household Longitudinal Study (UKHLS), which is gradually moving away from a single-mode, computer-assisted personal interviewing (CAPI) design towards a sequential Web-CAPI mixed-mode design (Lynn, 2017).

Sequential mixed-mode designs have several practical and methodological advantages over single-mode designs. First, such designs can lead to potential cost savings, particularly when the sequence starts with the least-expensive mode (Bianchi, Biffignandi, & Lynn, 2017; Wagner, Arrieta, Guyer, & Ofstedal, 2014). Second, such designs can improve coverage by allowing segments of the population that lack access to a particular mode (e.g. Web) to participate in the survey. One example is the German GESIS Panel, which uses a mix of Web and mail surveys in order to reach persons without internet access (Bosnjak et al., 2018). Third, sequential mixed-mode surveys can improve response rates and reduce the risk of nonresponse bias by reaching different kinds of respondents who possess different mode-specific preferences (De Leeuw, 2005; Roberts, Joye, & Stähli, 2016). All of these advantages have led to an increasing use of sequential mixed-mode designs and it is likely that such designs will continue be a mainstay of survey research for the foreseeable future.

However, sequential mixed-mode designs have an important drawback, which is the potential for differential measurement errors. Such errors arise when data are collected from different subsets of respondents through different modes which have inherently different measurement error properties. For example, self-administered modes are known for their tendency to elicit more honest responses and less social desirability bias for sensitive items than interviewer-administered modes (Tourangeau, Conrad, & Couper, 2013; De Leeuw, 1992; Cernat, Couper, & Ofstedal, 2016). Other types of measurement errors (e.g. acquiescent response style, extreme response style, among others; see Kieruj & Moors, 2010) have also been shown to vary between modes (De Leeuw & Hox, 2011; Aichholzer, 2013). Mixing multiple

Contact information: Alexandru Cernat, University of Manchester, Humanities, Bridgeford Street-2.13N, Manchester M13 9PL (email: alexandru.cernat@manchester.ac.uk)

modes that differentially influence respondents' answers produces at least two unwanted effects when the data are combined for analysis. First, it compromises the accuracy of comparisons of respondents who are interviewed in different modes, and second, it compromises the accuracy of comparisons involving other surveys which employ a single mode of data collection. The effects of mixing modes can be particularly problematic in longitudinal studies, where measures of change over time may reflect measurement effects rather than actual temporal changes in respondents' status (Cernat, 2015).

In this article, we examine the effects of mixing modes on measurement effects by considering multiple sources of measurement error which are analyzed simultaneously through a quasi-experimental design implemented in several waves of a longitudinal study. We use a novel model to estimate measurement error called multitrait-multierror (MTME, Cernat & Oberski, 2018; Cernat & Oberski, 2020). In our implementation of this approach we experiment with three types of measurement error: social desirability, acquiescence, and method effect. For each type of measurement error, we design an experiment to manipulate it and make it possible to estimate it. To investigate social desirability, we change the stem of the question offering either a positively- or a negatively-worded question (in order to manipulate the perceived social norm). For acquiescence, respondents are randomly presented with either an agree-disagree response scale or the reversed scale order: disagree-agree. Finally, for method effects we randomly allocate respondents to questions using either a 2-point scale or an 11-point scale. Thus, in this paper we define social desirability as the impact of a positively- or negatively-worded question stem, acquiescence as the impact of scale order (agree-disagree vs. disagree-agree), and method effect as the influence of the response scale (2 vs. 11 points).[1]

In the next section, we provide a review of the literature on measurement effects in mixed-mode designs and present the research questions. In the subsequent section we describe the methodology and data sources used in the study. Lastly, the study results and discussion are presented.

2 Literature Review

There is a large body of work on the effects of mixing modes on measurement error. Numerous studies have shown that different data collection modes can affect data quality in different ways (DeMaio, 1984; Tourangeau, Rips, & Rasinski, 2000; Heerwegh & Loosveldt, 2011; Ansolabehere & Schaffner, 2011; Hope, Campanelli, Nicolaas, Lynn, & Jäckle, 2014; McClendon, 1991; Holbrook, Krosnick, Moore, & Tourangeau, 2007; Dillman et al., 2009; Ye C. & Tourangeau, 2011; Smyth, Olson, & Kasabian, 2014; Nicolaas, Campanelli, Hope, Jäckle, & Lynn, 2015; Revilla, Saris, Loewe, & Ochoa, 2015). The most alarming conclusion drawn from this literature is that different modes can evoke different responses to the same questions by the same group of respondents. Mixing modes with different measurement properties can therefore give rise to so-called measurement effects (see e.g., Klausch, J.J., & Schouten, 2013). Two important contextual factors that explain differences in mode-specific measurement errors are the communication channel and the presence or absence of an interviewer. The communication channel refers to whether the survey is administered visually (e.g. Web), orally (e.g. computer-assisted telephone interviewing; CATI), or both (as is sometimes the case in CAPI surveys). It is thought that visual and oral communication channels differentially affect cognitive processes and memory capacity. For example, Krosnick and Alwin (1987) suggest that respondents may be more likely to choose response options listed first in a visual survey mode (primacy effect) as these options undergo deeper cognitive processing than later options. The authors also suggest that respondents may be more likely to choose response options presented last in an oral survey mode due to limits of working memory capacity. Modes which lack interviewer presence are known to produce a greater sense of privacy among respondents, inducing more candid answers and self-disclosure compared to interviewer-administered modes (Tourangeau & Yan, 2007).

The implications of these mode features for different types of measurement errors have been studied in single-mode comparisons (e.g. CAPI vs. Web) and, to a lesser extent, in mode design comparisons, which compare, for example, a single-mode (e.g. CAPI) design with a sequential mixed-mode (e.g. Web-CAPI) design. We review this literature in the context of the three types of measurement errors investigated in this paper: social desirability, acquiescence, and method effects.

2.1 Social desirability

Social desirability refers to the tendency for respondents to provide answers that present themselves in a more favorable light with respect to social and societal norms. Several studies have demonstrated that respondents are more likely to provide socially desirable answers to sensitive questions in interviewer-administrated modes than in self-administered modes (Tourangeau & Smith, 1996; De Leeuw, 2005). For example, De Leeuw (1992) showed in a meta-analysis that mail surveys produce less social desirability bias than telephone and face-to-face surveys. A more recent meta-analysis

[1] Measurement error can be defined and labelled in different ways depending on the discipline and method used. For example, what we call acquiescence is sometimes called acquiescence response style. Similarly, method effect can refer to a confounding of the number of response scale points and the use of labels or it can refer to different raters. Here, we explicitly define the measurement errors of interest in the introduction and method section in order to avoid any confusion.

by Tourangeau et al. (2013) showed that Web surveys produce less social desirability bias than interviewer modes. In sum, the literature suggests that self-administered modes (e.g. mail, Web) yield the least amount of social desirability, followed by CAPI, and then CATI (Bowling, 2005; Tourangeau & Yan, 2007; Heerwegh, 2009; Cernat et al., 2016) with interactive voice response (IVR) in-between Web and CATI (Kreuter, Presser, & Tourangeau, 2008).

In the context of mode design experiments, a few studies have shown that mixing a self-administered mode with an interviewer-administered mode can produce measurement effects with respect to social desirability. Vannieuwenhuyze, Loosveldt, and Molenberghs (2012), for example, report measurement effects in an experimental comparison involving a single-mode CAPI design with a sequential mail and CAPI design for questions about attitudes towards surveys. In an experimental mode design study of non-western immigrants in the Netherlands, Kappelhof and De Leeuw (2017) report an increase in measurement error variance effects for sensitive items when multiple sequential modes of data collection (Web-CATI-CAPI) are used in combination with interviewers who have a shared migration background with respondents, relative to a single-mode CAPI design.

2.2 Acquiescence

Acquiescence is a "weak form" of satisficing (Krosnick, 1991) that refers to the tendency for respondents to indiscriminately agree or answer "yes" to statements regardless of their content—presumably, because it is less cognitively demanding to agree than to disagree (Knowles & Condon, 1999). De Leeuw (1992) found that a self-administered (mail) questionnaire resulted in less acquiescence than interviewer-administered questionnaires. Similar results were found in earlier studies (e.g., Dillman & Mason, 1984; Tarnai & Dillman, 1992), but De Leeuw (2005) notes that the results are modest and not always consistent. Fricker, Galesic, Tourangeau, and Yan (2005) found no differences in acquiescence in an experimental comparison of Web and telephone interviewing. In an experimental comparison of CAPI and Web modes, Heerwegh (2009) reports slightly more acquiescence in the face-to-face mode, although the result was not statistically significant.

2.3 Method effect

A method effect refers to characteristics of the response scale that influence the way respondents answer questions (Andrews, 1984; Saris & Gallhofer, 2007). For example, the number of points in a response scale is a commonly-studied method effect (Revilla, Saris, & Krosnick, 2014). Method effects in the mixed-mode literature are often studied using multitrait-multimethod (MTMM) experiments (Campbell & Fiske, 1959; Jöreskog, 1970; Althauser & Heberlein, 1970; Alwin, 1974; Andrews, 1984). MTMM experiments involve repeating several questions ("traits") using several methods (e.g. different response scales, different response category labels) and using reliability and validity coefficients to assess the quality of the responses (e.g., Saris & Andrews, 1991). In a mixed-mode MTMM experiment conducted in the Netherlands component of the European Social Survey (ESS), Revilla (2010) found some differences in data quality (defined as the product of the squared reliability and validity coefficients) between CATI and two other modes (Web and CAPI) when the response options were experimentally varied (6 points, 8 points, 11 points, and open-ended). However, fewer differences were found when comparing the single-mode CAPI design with two mixed-mode designs involving all three modes. In another experimental study, Revilla (2012) compared CAPI and Web modes in two separate surveys in the Netherlands that implemented several response scale variations and found little difference in data quality (i.e. product of squared reliability and validity coefficients) between both modes. In a more recent study, Revilla et al. (2015) compared a single-mode CAPI design to a sequential mixed-mode (Web-CAPI) design in two ESS country surveys (UK and Estonia) with varying response scale labels and found similar estimates of data quality coefficients between both mode designs.

2.4 Study Limitations and Research Questions

Studying measurement effects in mixed-mode surveys faces several challenges. For example, one of the challenges faced in many of the above studies is separating selection effects from measurement effects. Mode selection effects are caused by differences between units with respect to their likelihood of completing the survey in a given mode. These mode-specific response propensities can affect the composition of respondents answering in a given mode and, in turn, confound the investigation of measurement effects. Although previous studies have tried to analyze measurement effects in mixed-mode surveys, it is rare that they account for selection effects in their analysis (e.g., Allum, Conrad, & Wenz, 2018; Heerwegh & Loosveldt, 2011; Gordoni, Schmidt, & Gordoni, 2012; Vannieuwenhuyze & Revilla, 2013). A further limitation of previous studies is that methods of assessing and correcting for measurement errors typically focus on a single type of measurement error (e.g. social desirability) independently of other measurement error types (e.g. acquiescence, method effects) (Couper, 2011). Assessing multiple errors simultaneously in a multivariate context has the potential to improve estimation of these errors as well as determine their relative contributions to the measurement accuracy for a given item. Such an approach is particularly useful for practitioners as it provides revealing information about how one might better allocate resources to minimize multiple sources of measurement error in surveys.

In this article, we address these research gaps by using a

quasi-experimental design implemented in multiple waves of a longitudinal mixed-mode (CAPI and Web) survey. Further, we use an innovative multitrait-multierror (MTME; Cernat & Oberski, 2018; Cernat & Oberski, 2020) modelling approach to assess the relative magnitude of multiple types of measurement error simultaneously while accounting for selection effects. With the results of this study we aim to provide researchers with a better understanding of the contributions of different measurement error sources that can arise in mixed-mode survey designs. The following research questions are addressed:

1. To what degree does using a mixed-mode (Web-CAPI) approach lead to different measurement errors compared to a single-mode (CAPI) approach?

2. To what degree do Web responses lead to different measurement errors compared to CAPI responses?

3 Data and Methods

3.1 UK Household Longitudinal Study – Innovation Panel (UKHLS-IP)

In this study we use the UK Household Longitudinal Study – Innovation Panel (UKHLS-IP; University of Essex, Institute for Social and Economic Research, 2018) to investigate the impact of mode and mode design on measurement error. The UKHLS-IP is a yearly household panel representative of the UK general population that is mainly used for methodological research[2]. The panel began collecting data in 2008 with 2,760 households being selected using random sampling with stratification and clustering (Jäckle, Al Baghal, Burton, & Lynn, 2018). Two refreshment samples were added in wave 4 (960 new addresses) and wave 7 (1,560 new addresses). The study started out as a CAPI survey but moved to a mixed-mode design in wave 5. In order to investigate the impact of mixing modes on data quality, one-third of the sample continues to be administered via single-mode CAPI while the remaining two-thirds are interviewed using a sequential Web-CAPI approach. Both groups also have a "mop-up" stage where remaining nonrespondents are recruited to participate via Web, CAPI, or telephone (for details, see Jäckle et al., 2018). We exclude the small number of cases that participated by telephone.

For this study, we focus on waves 7, 8, and 9. The conditional household-level response rates (conditional on being eligible) for the relevant waves were 78.5%, 82.7%, and 84.7%, respectively, while the conditional (on participating in the household survey) individual-level response rates were 82%, 85.4%, and 85.4%, respectively (Jäckle et al., 2018). In wave 7 the conditional individual-level response rates for the single mode (CAPI) design was 81.5% compared to 82.4% for the mixed-mode (Web-CAPI) one. In wave 8 these were both around 85% while in wave 9 they were 82.6% for single-mode versus 86.9% for the mixed-mode design (Jäckle et al., 2018).

3.2 Multitrait-Multierror (MTME) Modelling Approach

The MTME approach was recently developed to deal with some of the inherent limitations of the MTMM approach (Cernat & Oberski, 2018; Cernat & Oberski, 2020). The strength of the MTMM is the ability to estimate random error and method effects using a within-experimental design. However, the MTMM model does make a strong assumption regarding the absence of any other type of measurement error. This assumption may not hold in some cases, such as when measuring attitudes towards immigrants, where other factors, such as social desirability or acquiescence, might be present. In the MTME approach used here multiple potential sources of error are experimentally manipulated and modelled. In a sense the MTMM is a special type of MTME in which only the method is manipulated. For more details on designing MTME, the reader is referred to Cernat and Oberski (2018), and Cernat and Oberski (2020). Next, we describe how the MTME was implemented in the UKHLS-IP.

3.3 Experimental Design

In waves 7, 8 and 9 the UKHLS-IP implemented a MTME experiment that manipulated the wording and response scales of six questions in order to estimate: social desirability, acquiescence and method effects. It is these estimates of measurement error which are the focus of this paper.

The MTME experiment was implemented using six questions regarding attitudes towards immigrants (Table 1). The design of the MTME experiment starts from the decision regarding the types of systematic errors one wants to estimate. In this case, these were social desirability, acquiescence, and method effects. For each type of systematic error, two possible manipulations were implemented. To impact the direction of social desirability the wording of the question was either positive (e.g., we should allow more immigrants) or negative (e.g., we should allow fewer immigrants). To manipulate the direction of acquiescence either an agree-disagree format of the response scale or a reversed scale format (disagree-agree) was used. Finally, to estimate method effects either a 2-point scale or an 11-point scale was used. By combining these three dimensions, eight different ways of asking about attitudes towards immigration were developed for a given item/trait (Table 2).

The implementation of the MTME is similar to the one used in earlier MTMM split ballot experiments (Saris, Satorra, & Coenders, 2004; Saris & Gallhofer, 2007). Each

[2]Data can be freely downloaded from the UK Data Archive: https://www.ukdataservice.ac.uk/. Syntax used for the analysis can be found here: https://github.com/alex-cernat/MTME-MM

Table 1
Items measuring attitudes towards immigrants

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

Table 2
Wording variations used in the MTME experiment for one example item

Wording number	Social desirability	Number of scale points	Agree or Disagree	Item formulation (using trait 1 as an example)
W1	Higher	2	AD	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	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	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	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	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	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	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	The UK should allow fewer people of the same race or ethnic group as most British people to come and live here

respondent was randomly assigned to receive the six items twice, once at the beginning of the survey and once at the end. The order of the forms was randomized. Overall, the average time between the two sets of questions in the survey was 30 minutes.

The approach has some strengths and limitations. The strengths lie in the fact that it's implemented within an experimental design. If, for example, we find a difference for the groups that received the 2 point vs. 11 point scales we are certain that the response scale is the cause (the other factors and order were randomized). This is also true for the other two dimensions. The limitations stem from the fact that our conclusions are only as good as our experimental design. Because of this we are more confident in the estimation and interpretation of method and acquiescence, as these were investigated previously using similar approaches. We are less confident about our manipulation of social desirability as there is limited research on how to change this behaviour simply by changing the question. While our manipulation of the social norm is in line with the theoretical mechanism of social desirability, it will only work if respondents are indeed influenced by this change. Previous research has found limited support for this type of manipulation (Cernat & Oberski, 2018).

3.4 Model Estimation

To estimate the MTME model, we use Bayesian Structural Equation Modelling with non-informative priors[3] as imple-

[3]For most coefficients we used the priors set by default by Mplus (Asparouhov & Muthén., 2010). For the variance of the measure-

mented in Mplus 8.3[4]. A reduced form of the model (trait 1 only) is shown in Figure 1. One can see that trait 1 (T1) is measured using eight different wordings (W1-W8, based on Table 2). These are the observed variables (represented by squares) which measure an unobserved/latent variable T1 (represented by a circle). Additionally, there are the three types of systematic measurement error. These can be estimated either using an effect coding approach, where 0 is the average of the two conditions (which is unobserved), or a dummy coding approach, where 0 is the reference condition. In our model, we use both approaches. For social desirability (S) and acquiescence (A) we use an effect coding approach while for the method effect (M) we use a dummy coding approach (similar to using an MTMM-1; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003). The choice between dummy coding or effect coding shouldn't change the conclusions but it does change the interpretation of the latent variable (similar to what happens when we use different codings in a regression). Here we used dummy coding for the method factor for estimation reasons. Doing this makes it easier to estimate the model and code the coefficients[5].

In order to identify the latent variables, different constraints for the loadings are used[6]. These depend on the experimental design in Table 2. One can see that all of the observed variables in Figure 1 measure trait 1 so all of them have their loadings fixed to "+1". Similarly, wordings 3, 4, 7, and 8 are measured using the 11-point scale so they have "+1" in relationship with the method effect. For social desirability and acquiescence, we use "+1" or "-1" depending on the type of wording used. Finally, we restrict all of the means/thresholds of the observed variables and estimate the means of the latent variables. The full model also includes the latent variables T2-T6 as well as the observed variables measuring those questions (these variables are not included in Figure 1 for ease of reading).

The measurement error estimates from the MTME are used to compare the effect of mode (design) in UKHLS-IP. The following steps are used. We start by using the simple MTME model in each wave and compare the posterior distributions of the three systematic measurement errors by mode (CAPI vs. Web) and mode design (CAPI vs. Web-CAPI). Secondly, we develop regression models where control variables are used to explain the three types of systematic errors. Thirdly, we develop a model that explains the measurement error in MTME using both control variables and mode (design). A summary of the modelling approach is presented in Table 3, while the model is shown below:

$$Y = \beta_0 + \beta_1 \cdot \text{mode(design)} + \beta_2 \cdot \text{controls} + \epsilon \qquad (1)$$

where Y represents the three types of measurement error, β_0 is the intercept term, β_1 is the effect of the mode (design), and β_2 is the vector of coefficients for the control variables.

As mentioned in the introduction and literature review, one of the biggest challenges in researching mode effects is the confounding of mode selection and mode measurement effects. To partially account for this confounding, the following control variables are added to the model: age, gender, having a partner, being white British, living in rural area, having a degree, and being employed, using the internet daily, answering the survey by mobile device and having a long term illness. The effects of mode are investigated only after controlling for these potential confounders. The randomization of mode design (CAPI vs. Web-CAPI) gives extra strength to the separation of selection and measurement.

Table 3
Modelling approach to estimate effect of mode (design) on measurement error.

	Mode (design) comparison
Exploration	Compare means and variances using posterior distributions
Inference	a. Regress measurement error on control variables
	b. Regress measurement error on control variables and mode
	c. Investigate mode regression coefficient and change in R^2

4 Results

The MTME models, described in the methods section, were fitted separately for the three waves of the UKHLS-IP. Table 4 presents the samples sizes, means, and variances of the three measurement error types. While the means indicate how much the average response on a trait is changed by an experimental condition, the variance indicates how much the average response on that trait varies within an individual. The mean of the measurement errors can bias point estimates while the variance of the measurement errors can bias standard errors and multivariate analyses. Both can be estimated

ment error we set an inverse gamma prior of alpha = 2 and beta = 1. For the variance covariance matrix of the traits we used an Inverse-Wishart distribution with parameters: 12 and 10.

[4]We used four chains and 10,000 iterations to estimate the models. We also used a thinning coefficient of 100 and a burn in of 50%. We use a Potential Scale Reduction of 0.05 to ensure convergence. We additionally investigated the trace plots and posterior distributions for any potential signs of misspecification. We also report the fit indices of the models in the appendix.

[5]This is because adding another latent variable would have a different link function and it would be difficult to code the loadings appropriately.

[6]We assumed linear relationships. For the two-point scale items we used a probit link function.

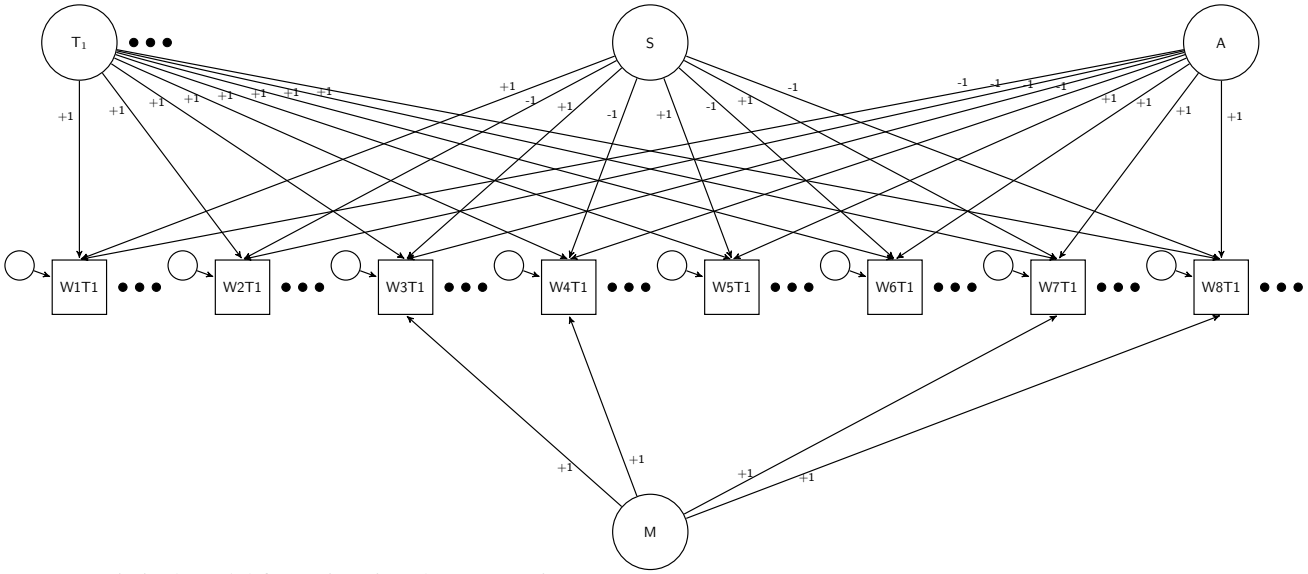

Figure 1. Statistical model for estimating the MTME in one wave

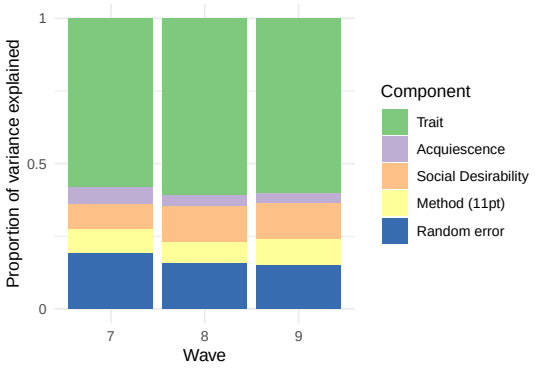

Figure 2. Variance decomposition by type of measurement error and by wave

using the MTME.

One can see that the three types of measurement error are present and they impact both the means and the variances in each of the three waves. Using each variance, the total variance of the six variables measuring attitudes towards immigration can be decomposed, as shown in Figure 2. The figure shows that trait (or "true score") represents around 60% of the total variance. Random error is the largest source of non-trait variance followed by social desirability, method, and acquiescence.

Based on these models, the posterior values of the three types of systematic errors are predicted by wave[7]. This makes it possible to explore differences in measurement error by mode and mode design. Looking at Table 5, there are very small differences by mode and mode design within each wave and statistic. The only important difference is in the means of social desirability, which is more negative in the mixed-mode design and for Web interviews compared to single mode and CAPI interviews. Because effect coding was used to estimate this measurement error, the way to interpret this finding is that changing the wording of the question from positive-to-negative shifts the average response more in the Web survey (and in the mixed-mode design) than in the CAPI (or single) mode design.

RQ1: To what degree does using a mixed-mode (Web-CAPI) approach lead to different measurement errors compared to a single-mode (CAPI) approach? Next, we investigate the differences in measurement error by mode design using the aforementioned regression approach. The regression slopes of the mode design on measurement error show the impact on their mean while the R^2 indicates how much of the variation in measurement error is explained by the mode design (after taking into account the control variables). Similar to what was observed in the descriptive analysis, the main consistent difference is in social desirability (Table 6).

The mixed-mode respondents have more extreme levels of social desirability compared to respondents allocated to the single-mode design. The mode design explains around 1% of the variation in social desirability, indicating that mode design is only a small part of the mechanisms behind this type of measurement error. We also observe no differences in the acquiescence and method effect between mode designs at any point in time.

[7]The use of posteriors may underestimate the uncertainty around the coefficients. They are used only for exploratory purposes. The use of regression in the next section takes into account uncertainty in the estimation of the measurement errors.

Table 4
Mean and variances for the three types of measurement error by wave

	Sample	Means			Variances		
Wave	size	Soc. des.	Method[a]	Acq.	Soc. des.	Method	Acq.
7	2,314	−0.18	5.13	0.25	0.29	0.86	0.42
8	2,246	−0.13	4.94	0.16	0.40	0.75	0.60
9	2,154	−0.33	5.03	0.26	0.98	0.88	0.44

[a] The interpretation of the mean of the method is indeed a little more difficult. If there was no effect of the method we would expect a mean of the method to be 5.5 (11/2). If the value is smaller then there is a tendency of respondents being less extreme when using the 11-point scale compared to the 2-point scale.

Table 5
Descriptive statistics of systematic errors by mode (design) and wave

			Means			Standard deviation		
		Wave	Soc. des.	Method	Acq.	Soc. des.	Method	Acq.
Mode design	Single	7	-0.31	5.14	0.25	0.70	0.81	0.55
	Mixed	7	-0.36	5.13	0.26	0.69	0.81	0.55
	Single	8	-0.10	4.97	0.17	0.47	0.73	0.63
	Mixed	8	-0.14	4.94	0.17	0.47	0.75	0.63
	Single	9	-0.16	5.05	0.25	0.63	0.77	0.52
	Mixed	9	-0.22	5.03	0.23	0.62	0.75	0.52
Mode	CAPI	7	-0.32	5.14	0.25	0.70	0.81	0.55
	Web	7	-0.39	5.12	0.26	0.69	0.81	0.55
	CAPI	8	-0.10	4.96	0.16	0.47	0.74	0.63
	Web	8	-0.15	4.96	0.18	0.47	0.74	0.64
	CAPI	9	-0.18	5.04	0.24	0.63	0.77	0.52
	Web	9	-0.23	5.03	0.24	0.62	0.75	0.52

Table 6
Regression coefficient and R^2 of mixed mode vs. single mode by wave (with controls).

Wave	ME	Est	Post SD	Lower C.I.	Upper C.I.	R^2 extra
7	Social desirability	−0.30*	0.10	−0.51	−0.10	0.0
	Method	−0.10	0.09	−0.27	0.08	0.3
	Acquiescence	0.01	0.06	−0.10	0.12	0.2
8	Social desirability	−0.68*	0.15	−0.99	−0.38	1.2
	Method	−0.08	0.09	−0.27	0.10	0.3
	Acquiescence	0.01	0.10	−0.19	0.21	0.2
9	Social desirability[a]	−0.50*	0.18	−0.85	−0.16	1.4
	Method[a]	0.03	0.10	−0.18	0.23	0.2
	Acquiescence[a]	−0.08	0.09	−0.25	0.11	0.8

[a] PSR indicates possible misspecification in full model
* Coefficients do not include zero in the credible interval.

RQ2: To what degree do Web responses lead to different measurement errors compared to CAPI responses? Lastly, we make a direct comparison of Web and CAPI responses in the three waves (Table 7). The results are similar to those seen previously. The main difference between modes is evident for social desirability. As was observed in Table 5, the expected mean for CAPI respondents on the social desirability variable is lower than for the Web respondents. This is consistent in all three waves with an R^2 ranging from around 1% to approximately 3%. No differences in acquiescence and method effect by mode of interview are present.

5 Discussion

This paper investigated the impact of mode and mode design on measurement error using a combination of experimental designs and statistical modelling. We leveraged an experimental design that randomly allocated respondents to either a single-mode CAPI survey or a sequential mixed-mode Web-CAPI design, as well as the implementation of a multitrait-multierror model with three types of systematic measurement errors: social desirability, acquiescence, and method effect. The experimental mode design together with the control variables gives us a strong design to separate the confounding of measurement and selection. At the same time, the MTME enables us to estimate multiple types of measurement error simultaneously. Finally, the use of three waves of longitudinal data allowed us to validate the findings and assess how stable they are over time.

Overall, we find small differences in measurement error across mode (designs). The descriptive analysis showed that the variance of the systematic errors is similar across mode (designs) while some differences were present in the means for social desirability. These findings were supported using regression models. Social desirability was systematically different by mode (design). Although the results show that there are mode differences in measures of social desirability, this explains only a small amount of variance. Additionally, we found no mode (design) differences with respect to method effects and acquiescence, which goes against some of the previous research in this area (e.g., De Leeuw, 1992; Revilla, 2010), but is consistent with other studies (Fricker et al., 2005; Revilla, 2012; Revilla et al., 2015).

The most surprising finding was the direction of the mode effect on social desirability. While most of the research in this area has shown that social desirability effects are more prominent in interviewer modes (e.g. face-to-face) compared to self-administered modes (such as Web), we find that changing the wording of the item in question has a larger impact on the mean of the observed responses in Web than in CAPI. A post-hoc explanation could be that wording changes are more salient in self-administered modes and that the experimental manipulation does not only encompass social-desirability, but also other types of measurement error.

As is the case with all research studies, this one has some limitations. The first limitation is our approach to estimating differences in measurement errors by mode makes the assumption that the measurement model is the same for the two mode groups. A way to free this assumption could be to run a multi-group model by mode (design). Unfortunately, this approach does not work with the available data due to sample size limitations. Additionally, it would be ideal to use such an approach in different countries, alternative mode designs, and different manipulations of the MTME. Investigating such variations is a topic for future work.

That being said, this paper contributes to the mixed-mode literature through the use of experimental designs both in dealing with the confounding of measurement and selection effects in mode (designs) as well as measurement confounding (i.e. looking at one measurement error at a time and ignoring the influence of the others) which is common in the survey literature. The research shows no differences between a particular single-mode (CAPI) and mixed-mode (Web-CAPI) design or between Web and CAPI modes with respect to acquiescence and method effect (number of response scale points). Although a different mixed-mode design (for example, including telephone) could lead to different results, the results at hand are reassuring for survey practice where mixed-mode designs involving Web and CAPI are becoming more common. On the other hand, practitioners should be aware of the potential that responses collected in a self-administered mode are more likely to be influenced by the wording of the question stem than in an interviewer mode, which may affect estimates of social desirability bias in multitrait experiments.

Acknowledgments

This paper is partially financed by the National Centre for Research Methods/Economic and Social Research Council grant [R121711]. We would like to thank Daniel Oberski for the initial work in developing the MTME approach and the original statistical models. We would also like to thank the Understanding Society Innovation Panel team for helping us implement the MTME design and for collecting the data free of charge.

References

Aichholzer, J. (2013). Intra-individual variation of extreme response style in mixed-mode panel studies. *Social Science Research*, *42*(3), 957–970.

Allum, N., Conrad, F., & Wenz, A. (2018). Consequences of midstream mode switching in a panel survey. *Survey Research Methods*, *12*(1), 43–58.

Althauser, R. & Heberlein, T. (1970). Validity and the multitrait-multimethod matrix. In E. Borgatta & G. Bohrnstedt (Eds.), *Sociological methodology*. San Francisco: Jossey-Bass.

Table 7
Regression coefficient and R^2 of CAPI vs Web by wave (with control variables).

Wave	ME	Est	Post SD	Lower C.I.	Upper C.I.	R^2 extra
	Social desirability	0.20*	0.05	0.11	0.30	3.1
7	Method	0.04	0.04	−0.05	0.13	0.3
	Acquiescence	−0.01	0.04	−0.09	0.07	0.2
	Social desirability	0.63*	0.17	0.31	0.97	1.1
8	Method	0.02	0.10	−0.18	0.23	0.3
	Acquiescence	0.04	0.11	−0.18	0.26	0.2
	Social desirability[a]	0.70*	0.19	0.34	1.07	1.2
9	Method[a]	−0.04	0.11	−0.26	0.16	0.3
	Acquiescence[a]	0.09	0.10	−0.10	0.28	0.7

[a] PSR indicates possible misspecification in full model
* Coefficients do not include zero in the credible interval.

Alwin, D. (1974). Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In H. Costner (Ed.), *Sociological methodology 1973-1974*. San Francisco: Jossey-Bass.

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

Ansolabehere, S. & Schaffner, B. (2011). Does survey mode still matter? findings from a 2010 multi-mode comparison. *Political Analysis*, *22*, 285–303.

Asparouhov, T. & Muthén., B. (2010). Bayesian analysis using Mplus. *Technical Implementation Mplus*.

Bianchi, A., Biffignandi, S., & Lynn, P. (2017). Web-face-to-face mixed-mode design in a longitudinal survey: Effects on participation rates, sample composition, and costs. *Journal of Official Statistics*, *33*, 385–408.

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A., & Weyandt, K. W. (2018). Establishing an open probability-based mixed-mode panel of the general population in germany: The GESIS panel. *Social Science Computer Review*, *36*(1), 103–115.

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, *27*(3), 281–291.

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

Cernat, A. (2015). The impact of mode design on measurement errors and estimates of individual change. *Survey Research Methods*, *9*(2), 83–99.

Cernat, A., Couper, M., & Ofstedal, M. (2016). Estimation of mode effects in the health and retirement study using measurement models. *Journal of Survey Statistics and Methodology*, *4*(4), 501–524.

Cernat, A. & Oberski, D. (2018). Estimating stochastic survey response errors using the multitrait-multierror model. National Centre for Research Methods, NCRM Working Paper. Retrieved from http://eprints.ncrm.ac.uk/4156/

Cernat, A. & Oberski, D. L. (2020). Extending the within-persons experimental design: The multitrait-multierror (MTME) approach. *Experimental methods in survey research*, New York: John Wiley & Sons.

Couper, M. (2011). The future of modes of data collection. *Public Opinion Quarterly*, *75*(5), 889–908.

De Leeuw, E. (1992). *Data quality in mail, telephone and face to face surveys*. Amsterdam: TT-Publ.

De Leeuw, E. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*, 233–255.

De Leeuw, E. (2018). Mixed-mode: Past, present, and future. *Survey Research Methods*, *12*(2), 75–89.

De Leeuw, E. & Hox, J. (2011). Internet surveys as part of a mixed-mode design. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social behavioral research and the internet*. New York: Taylor and Francis.

DeMaio, T. (1984). Social desirability and survey measurement: A review. In C. G. Turner & E. Martia (Eds.), *Surveying subjective phenomena*. New York: Russell Sage Foundation.

Dillman, D. & Mason, R. (1984). The influence of survey methods on question response. Paper Presented at the Annual Meeting of the American Association for Public Opinion Research, Buck Hill Falls, Pennsylvania.

Dillman, D., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. (2009). Response rate and

measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the internet. *Social Science Research*, *38*, 1–18.

Eid, M., Lischetzke, T., Nussbeck, F., & Trierweiler, L. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C (M-1) model. *Psychological Methods*, *8*(1), 38–60.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, *69*(3), 370–392.

Gordoni, G., Schmidt, P., & Gordoni, Y. (2012). Measurement invariance across face-to-face and telephone modes: The case of minority-status collectivistic oriented groups. *International Journal of Public Opinion Research*, *24*, 185–207.

Heerwegh, D. (2009). Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research*, *21*(1), 111–121.

Heerwegh, D. & Loosveldt, G. (2011). Assessing mode effects in a national crime victimization survey using structural equation models: Social desirability bias and acquiescence. *Journal of Official Statistics*, *27*, 49–63.

Holbrook, A., Krosnick, J., Moore, D., & Tourangeau, R. (2007). Response order effects in dichotomous categorical questions presented orally: The impact of question and respondent attributes. *Public Opinion Quarterly*, *71*, 325–348.

Hope, S., Campanelli, P., Nicolaas, G., Lynn, P., & Jäckle, A. (2014). The role of the interviewer in producing mode effects: Results from a mixed modes experiment comparing faceto-face, telephone and web administration, no. 2014-20. ISER Working Paper Series. Retrieved from https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2014-20.pdf

Jäckle, A., Al Baghal, T., Burton, J., & Lynn, P. (2018). Understanding society the UK household longitudinal study innovation panel, waves 1-10, user manual, 1–165. University of Essex, Colchester: Institute for Social and Economic Research. Retrieved from https://www.understandingsociety.ac.uk/sites/default/files/downloads/documentation/innovation-panel/%20user-guides/ip_user_guide.pdf

Jöreskog, K. (1970). A general method for estimating a linear structural equation system. ETS Research Report Series, 2, i-41.

Kappelhof, J. & De Leeuw, E. (2017). Estimating the impact of measurement differences introduced by efforts to reach a balanced response among non-western minorities. *Sociological Methods & Research*.

Kieruj, N. & Moors, G. (2010). Variations in response style behavior by response scale format in attitude research. *International Journal of Public Opinion Research*, *22*(3), 320–342.

Klausch, H., T., J.J., & Schouten, B. (2013). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research*, *42*, 227–263.

Knowles, E. & Condon, C. (1999). Why people say "yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, *77*(2), 379–386.

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*, 847–865.

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

Krosnick, J. & Alwin, D. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, *51*(2), 201–219.

Lynn, P. (2017). Pushing household panel survey participant from CAPI to web. Presented at the 28th International Workshop on Household Survey Nonresponse, Utrecht, August/September.

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

Nicolaas, G., Campanelli, P., Hope, S., Jäckle, A., & Lynn, P. (2015). Revisiting "yes/no" versus "check all that apply": Results from a mixed modes experiment. *Survey Research Methods*, *9*(3), 189–204.

Revilla, M. (2010). Quality in unimode and mixed-mode designs: A multitrait-multimethod approach. *Survey Research Methods*, *4*(3), 151–164.

Revilla, M. (2012). Impact of the mode of data collection on the quality of answers to survey questions depending on respondent characteristics. *Bulletin de Methodologie Sociologique*, *116*, 44–60.

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

Revilla, M., Saris, W., Loewe, G., & Ochoa, C. (2015). Can a non-probabilistic online panel achieve question quality similar to that of the european social survey? *International Journal of Market Research*, *57*(3), 395–412.

Roberts, C., Joye, D., & Stähli, M. (2016). Mixing modes of data collection in swiss social surveys: Methodological report of the LIVES-FORS mixed mode experiment. LIVES Working Paper 2016/48, Swiss National Centre of Competence in Research.

Saris, W. & Andrews, F. (1991). Evaluation of measurement instruments using a structural modeling approach. In

P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys*. New York: John Wiley.

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

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.

Smyth, J., Olson, K., & Kasabian, A. (2014). The effect of answering in a preferred versus a non-preferred survey mode on measurement. *Survey Research Methods*, *8*(3), 137–152.

Tarnai, J. & Dillman, D. (1992). Questionnaire context as a source of response differences in mail versus telephone surveys. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research*. New York: Springer-Verlag.

Tourangeau, R., Conrad, F., & Couper, M. (2013). *The science of web surveys*. New York: Oxford University.

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

Tourangeau, R. & Smith, T. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, *60*(2), 275–304.

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883.

U.S. Census Bureau. (2014). American community survey: Design and methodology. Technical Report, Version 2.0. Retrieved from http : / / www2 . census . gov / programs - surveys / acs / methodology / design _ and _ methodology / acs _ design _ methodology % 20report _ 2014.pdf

University of Essex, Institute for Social and Economic Research. (2018). Understanding society: Innovation panel, waves 1-10, 2008-2017, [data collection] 9th edition. UK Data Service, SN: 6849. Retrieved from http://doi.org/10.5255/UKDA-SN-6849-10

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2012). A method to evaluate mode effects on the mean and variance of a continuous variable in mixed-mode surveys. *International Statistical Review*, *80*(2), 306–322.

Vannieuwenhuyze, J. & Revilla, M. (2013). Evaluating relative mode effects on data quality in mixed-mode surveys. *Survey Research Methods*, *7*(3), 157–168.

Wagner, J., Arrieta, J., Guyer, H., & Ofstedal, M. (2014). Does sequence matter in multimode surveys: Results from an experiment. *Field Methods*, *26*, 141–155.

Ye C., J., Fulton & Tourangeau, R. (2011). More positive or more extreme? a meta-analysis of mode differences in response choice. *Public Opinion Quarterly*, *75*, 349–365.

Appendix
Tables

Table A1
Number of respondents answering in each mode (design) by wave

	Outcome	Wave 7	Wave 8	Wave 9
Mode design	Single mode	805	1244	744
	Mixed mode	1608	1134	1493
	Missing	509	544	685
Mode	Web	752	799	1123
	Face to face	1581	1439	1020
	Missing	589	684	779

Table A2
Model fit indicators

Model	Wave	Sample size	Lower χ^{2*}	Upper χ^{2*}	p-value**
Baseline with controls	7	2094	119.6	467.0	0.00
	8	1742	15.9	351.3	0.02
	9	1634	−15.6	328.1	0.04
Mixed mode control	7	2094	121.0	466.5	0.00
	8	1742	8.9	350.0	0.02
	9	1634	−23.8	319.5	0.05
Mode control	7	2090	127.1	479.3	0.00
	8	1725	13.1	347.0	0.02
	9	1634	−15.6	328.1	0.04

* 95% Confidence interval for the difference between the observed and the replicated χ^2
** Posterior Predictive p-value