The impact of mixing modes on reliability in longitudinal studies

Alexandru Cernat
Institute for Social and Economic Research
University of Essex

No. 2013-09
July 2013

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

Non-Technical Summary

Questions in surveys can be asked in different ways, from face-to-face to self-administration on the internet. Furthermore, these different ways of asking can be mixed both for the same respondent and across individuals. These decisions influence the quality of the data that users can subsequently analyse.

In this study I compare a design that applies questionnaires face-to-face to one that uses a combination of telephone and face-to-face. The design of the data is used to see to what degree repeating the same questions in these different designs leads to the same responses (reliability).

Results show that the two designs, single mode face-to-face and multi mode telephone/face-to-face, lead to equally reliable and stable data for the 33 variables analysed. Speculations are made that selection and systematic errors may be more important factors for differences between different ways of administering questionnaires.

The impact of mixing modes on reliability in longitudinal studies

Alexandru Cernat*
Institute for Social and Economic Research, University of Essex

*Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK (email: acerna@essex.ac.uk)

Abstract

Mixed mode designs are increasingly important in surveys and large longitudinal studies are progressively moving to or considering such a design. In this context our knowledge regarding the impact of mixing modes on data quality indicators in longitudinal studies is sparse. This study tries to ameliorate this situation by taking advantage of a quasi-experimental design in a longitudinal survey. Using models that estimate reliability for repeated measures, quasi-simplex models, 33 variables are analysed by comparing a single mode CAPI design to a sequential CATI-CAPI design. Results show no differences in reliabilities and stabilities across mixed modes either in the wave when the switch was made or in subsequent waves. Implications and limitations are discussed.

Key words: longitudinal survey, mixed mode survey, CAPI, CATI, reliability, quasi-simplex models, latent Markov chains.

JEL Codes: C81, C83

Acknowledgements: I would like to thank all the people that helped with this paper: Peter Lynn, Nick Allum, Peter Lugtig, Noah Uhrig, Jakob Petersen

1 Introduction

Surveys are a mainstay institution in modern society, being essential for politics, policy, academic and marketing research and mass-media. In this context, the dropping response rates are threatening external validity (de Leeuw and de Heer, 2002). In parallel, the economic downturn adds pressure on survey agencies to decrease the overall price of surveys. In response to this data collection agencies are looking to both old solutions, such as increasing the number of contact attempts, and to newer ones, such as mixing modes, tailoring designs (Dillman et al., 2008) or using social media (Groves, 2011).

Mixing modes is one of the most important solutions considered in this context as it potentially leads to decreased overall cost without threatening data quality. This is done by maximizing responses in cheaper modes while using the more expensive modes in order to interview the hard to contact or unwilling respondents. In addition, the modes combined in this kind of design may lead to different coverage and non-response biases that can compensate each other. But, although mixing modes offers a good theoretical solution to saving costs its impact on data quality is still marred with unknowns.

More recently, longitudinal studies are also considering mixing modes as a solution to saving costs. The British Cohort Studies (e.g., National Child Development Study) and Understanding Society are such examples (Couper, 2012), the former already collecting data using mixed modes while latter is considering it. Unfortunately there are still many unknowns regarding mixing modes in this context. One important risk for this survey design in longitudinal studies is the potential increase of long-term attrition (Lynn, 2012) and its subsequent impact both on external validity and power. Additionally, mixing modes can lead to (different) measurement bias. This may, in turn, cause measurement inequivalence compared both with previous waves and with different modes.

Another aspect of the mixed mode design that has been relatively ignored in the literature so far and is especially important in longitudinal studies is the impact on reliability. Although cross-sectional mode comparisons usually concentrate on bias this represents only a part of the measurement issue. Different reliabilities in mixed-modes may be a threat to the longitudinal comparability of panel studies, confounding true change with change in random errors. More generally, reliability is an essential component of overall validity (Lord and Novick, 1968) as the random errors attenuate the relationship with other criterion variables. Empirically distinguishing between reliability and validity would help us understand the processes resulting from mixing modes and find possible solutions to minimize the differences across mode designs.

The present paper aims to tackle part of these issues by analysing the impact of mixing modes on data quality in a longitudinal study using a quasi-experimental design. The Understanding Society Innovation Panel (USIP), a national representative longitudinal study aimed at conducting methodological experiments, included a mixed mode design in its second wave. Here a sequential mixed mode design using Computer Assisted Telephone Interview (CATI) - Computer Assisted Personal Interview (CAPI) was randomly allocated and compared to a CAPI single mode design. This context will give the opportunity to use models that take advantage of the longitudinal character of the data (i.e., Quasi-Markov Simplex Models (QMSM) and Latent Markov Chains (LMC)) in order to compare the reliability of the two mode designs. The two models define reliability as the proportion of variance of the observed items that is due to the true score, as opposed to random error, and is consistent with Classical Test Theory (Lord and Novick, 1968, CTT).

2 Background

2.1 The impact of mixing modes and reliability

Mixing modes in surveys is becoming an increasingly important topic as it may offer some of the methodological solutions needed in the present context. There are three main reasons why this design is attractive. Firstly, it can decrease coverage error if the different modes reach different populations. A similar effect is obtained by minimizing non-response error. This is done by starting with a cheaper mode and sequentially using the more expensive modes to convert the hard to contact or unwilling respondents (De Leeuw, 2005). This would result in more representative samples as people who would not be reached by a certain mode would be included in the survey by using the other one. By using a combination of modes it is also believed that we could reduce costs by interviewing as many people as possible with the cheaper modes.

Modes can be mixed at various stages of the survey in order to achieve different goals. De Leeuw (2005) highlights three essential stages when these can be implemented: recruitment, response and follow-up. By combining these phases with the different types of modes results in a wide variety of possible approaches that try to minimize costs, nonresponse and measurement bias. The most important phase for our purposes is the second one (i.e., response), the mode used in this stage leading to the most important measurement effects. Therefore, the present article concentrates on this aspect of mixed modes.

Although mixing modes is attractive for the reasons listed above this approach also introduces heterogeneity that can affect data quality and substantive results. A large number of studies have tried to compare the modes and explain the differences found between them but there are still many unknowns regarding the mechanisms through which these appear. Tourangeau et al. (2000) provide one possible framework for understanding these. They propose three main psychological mechanisms through which modes lead to different responses. The first one is impersonality and it is affected by the respondents' perceived risk of exposing themselves due to the presence of others. The second dimension is perceived legitimacy of the survey and of the interviewer. The last one is the cognitive burden that each mode inflicts on the respondent. These can have an impact on any of the four cognitive stages of the response process: comprehension, retrieval, making judgements and selection of a response (Tourangeau et al., 2000, p. 7). This framework will be used in order to understand the mechanisms that may lead to differences across mode design.

When evaluating the impact of mixing modes on measurement usually the analysis concentrates either on missing data or on response styles such as acquiescence, primacy/recency or non-differentiation (Roberts, 2007; Betts and Lound, 2010; Dex and Gumy, 2011, for an overview). Although response styles are important reliability is an aspect that is often ignored in the mixed mode literature. As mentioned in the introduction, reliability is an important part of overall validity of the measurement (Lord and Novick, 1968) as it can attenuate the relationship with other (criterion) variables. Thus, differences in covariances that can be found between mode designs may be due to different percentage of random error rather than bias per se. This may prove to be an important distinction if we aim to understand the mechanisms that are leading to biased responses in different mode designs.

Furthermore, reliability is essential for longitudinal surveys. If different mode designs are implemented during the lifetime of a panel study the different reliability coefficients across modes can lead to artificial increase or decrease of change. These, in turn, having effects on the substantive results provided by the data. Understanding the level of reliability and the differences between modes on this indicator would help us comprehend to what degree this is an important issue.

Considering the present theoretical framework the reliability of the data in longitudinal studies can be influenced by four distinct factors. The first one of these is driven by the fact that cheaper modes are usually used in the mixed mode design. The mechanism is the direct effect of collecting data in an alternative mode that increases the respondent burden and decreases motivation. An example of this is CATI, which uses only the auditory communication channel, this increasing the burden on the respondent (De Leeuw, 2005). Telephone interviews are also on average shorter compared to CAPI, this causing further cognitive burden. In addition, the distance to the interviewer, both physical and social, means that the respondent is less invested in the completion of the questionnaire, this leading to lower quality data and more drop-offs. All these effects can lead to the increase of mistakes when responding to questionnaires using CATI and, therefore, to different degrees of reliability across modes.

The second mechanism is due to the different systematic errors specific to each mode. In order to illustrate the process I will use recency (e.g., McClendon, 1991, the tendency to select the last category) and primacy (e.g., Krosnick and Alwin, 1987, tendency to select the first category) response styles as examples. We know that we can expect higher degrees of primacy in visual modes, such as CAPI with showcards, while recency is stronger in the modes that use only the audio channel, such as CATI (Groves and Kahn, 1979; McClendon, 1991; Holbrook et al., 2007). If the mode specific effects are stable in time then models that estimate reliability, such as the quasi-simplex models, would overestimate reliability by including the systematic bias in the true score. Switching the mode, and changing the response style that is linked with it, leads to the movement of the variance due to the response style from the true score to the random error part of the model (i.e., the disturbance of the true score). Therefore, in the wave when the mode is switched we expect lower reliability as the mode specific systematic error is separated from the true score. This is true for all response styles that are mode specific and stable in time. This is also true for all the systematic mode specific effects caused by satisficing (Krosnick, 1991; Krosnick et al., 1996). In this framework respondents that have lost the motivation to complete the questionnaire in an *optimized* way will choose to bypass some of the mental steps needed in the response process. Satisficing can be either weak, such as selection of first category or acquiescence, or strong, like social desirability or the random coin flip (Krosnick, 1991). Thus, if the modes lead to a stable satisficing process then we would expect a decrease in reliability proportional with the size of the mode specific response bias and the proportion of the sample that responds using the new mode.

The third mechanism through which reliability can be influenced by mixing modes in longitudinal studies is panel conditioning. This is the process through which subjects change their responses because of the exposure to repeated measurements in time. This results in increase reliability and stability of items and decrease of item nonresponse (e.g., Jagodzinski et al., 1987; Sturgis et al., 2009; Chang and Krosnick, 2009). Therefore, changing the mode of interview may lead to the decrease of this effect if the mode change leads to the practice of a different cognitive task. If this is true then the reliability for the mixed mode design should be smaller in subsequent waves (Dillman, 2009).

The last factor leading to lower reliability in a mixed-mode design is the overall increase of the survey complexity. This, in turn, can lead to increase in errors both during the fieldwork and during the processing of the data. If this is true then we would expect differences in reliability between the two mode designs especially in the waves when we have multiple modes and less so in subsequent waves. Table 1 summarizes the possible effects of mixing modes on reliability in panel data compared to a single mode design.

Table 1: Mixed modes effects on reliability in a panel study

Cause	Mechanism	Waves affected
Direct	Burden and motivation	When modes are mixed
Mode switch	Change of systematic bias	When modes are mixed
Panel conditioning	Changing cognitive tasks	When modes are mixed and subsequent waves
Survey complexity	Errors in data collection and processing	When modes are mixed

So far relatively few studies have concentrated on quality indicators like reliability or validity in the mixed modes literature (e.g., Jäckle et al., 2006; Chang and Krosnick, 2009; Révilla, 2010, 2011; Vannieuwenhuyze and Révilla, in press). For example, Révilla (2010) has found small mean differences in the reliabilities of items measuring dimensions such as political trust, social trust or satisfaction using an MTMM design. The highest difference was found between a CATI and Computer Assisted Web Interview mode in the political trust model. Unfortunately these results are confounded with selection effects. A similar approach was applied using an instrumental variable that aimed to bypass this issue (Vannieuwenhuyze and Révilla, in press). Although some methodological limitations remain initial results show small to medium measurement effects and relatively large selection effects. The present paper will contribute to this literature by adding a new analytical model that takes advantage of the longitudinal data and offers an estimation of reliability.

2.2 Reliability in panel data

In order to evaluate the effect of the mixed mode design on the data quality I will concentrate on evaluating the impact on reliability. Using CTT we can define the reliability as the percentage of variance of the observed variable that is due to the true score as opposed to variance caused by random error (Lord and Novick, 1968). There are a number of models that aim to separate random measurement and true scores such as Multitrait-Multimethod

(Campbell and Fiske, 1959), Confirmatory Factor Analysis (Bollen, 1989, CFA) or the Quasi-Markov Simplex Model (Heise, 1969; Wiley and Wiley, 1970; Alwin, 2007).

Considering the characteristics of our data, four waves of panel data, I concentrate on the strand of literature that tries to explain reliability using repeated measures as opposed to multiple items (Alwin, 2007). A first attempt of assessing reliability using these kinds of measures was made by Lord and Novick (1968) who highlighted that by using two *parallel measures* we could estimate reliability. This term refers to measures that have equal true scores and equal variances of the random errors. If this is true then the correlation between the two measures is a correct estimation of reliability. But, as the authors themselves highlight (Lord and Novick, 1968, p. 134), this approach assumes the absence of memory, practice, fatigue or change in true scores. Especially the latter and the former make this estimation of reliability unfeasible for most social science applications.

In order to overcome the assumptions of the test-retest approach a series of models that take into account the change in time of the true scores were put forward. They usually assume an autoregressive change in time where the true score T_i is influenced only by T_{i-1} and no other previous measures. As a result, these models need at least three waves to be identified. In addition, they still need to make the assumption of equal variance of random error in order to be estimated (Wiley and Wiley, 1970; van de Pol and Langeheine, 1990). On the other hand they offer two important advantages (Alwin, 2007, p. 103). Firstly, they are able to separate random error from the specific variance of the true score. Secondly, under certain conditions, they can rule out systematic error as long as it is not stable in time.

In the next subsections I will present two such models. Although they are conceptually similar, imposing comparable assumptions and leading to estimates of reliability, they are developed from distinct statistical traditions and for different types of variables. As a result, QMSM can be used for continuous and ordinal variables by considering the true score continuous while the LMC model has been developed to deal with categorical variables and views the true scores as discrete.

2.2.1 Quasi-Markov Simplex Model

The QMSM is composed of two parts. The first one, the measurement component, is based on CTT, and assumes that the observed score A_i is caused by a true score, T_i, and random measurement error, ϵ_i. The impact of the true score on the observed variable is estimated with a regression slope λ_{ii}. The relationships in the case of a four waves model are:

$$A_1 = \lambda_{11}T_1 + \epsilon_1 \tag{1}$$

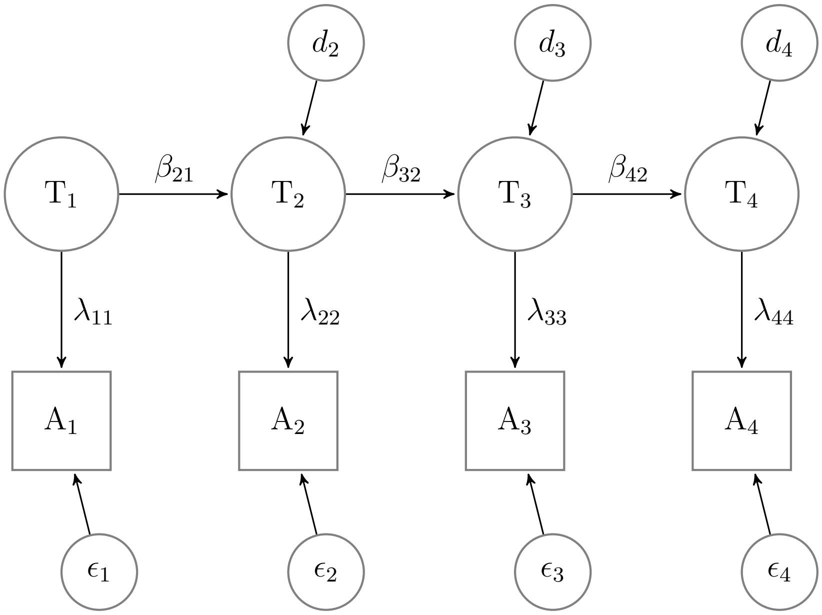
$$A_2 = \lambda_{22}T_2 + \epsilon_2 \tag{2}$$

$$A_3 = \lambda_{33}T_3 + \epsilon_3 \tag{3}$$

$$A_4 = \lambda_{44}T_4 + \epsilon_4 \tag{4}$$

In addition to the measurement part the model includes a structural dimension which models the relationships between the true scores. As a result of the auto-regressive (simplex)

Figure 1: Quasi–Markov Simplex Model for four waves



change in time of the true scores we have the following equations:

$$T_2 = \beta_{21}T_1 + d_2 \tag{5}$$

$$T_3 = \beta_{32}T_2 + d_3 \tag{6}$$

$$T_4 = \beta_{43}T_3 + d_4 \tag{7}$$

Where $\beta_{i,i-1}$ is the regression slope of T_{i-1} on T_i and d_i is the disturbance term. The former can be interpreted as stability in time of the true score while the latter can also be interpreted as the specific variance of the true score at each wave. The model can be seen in Figure 1.

In order to identify the model we need to make two assumptions. The first one constrains unstandardized λ_{ii} to be equal to 1:

$$\lambda_{11} = \lambda_{22} = \lambda_{33} = \lambda_{44} = 1 \tag{8}$$

In addition, I constrain the variance of the random errors, θ_i, to be equal in time (Wiley and Wiley, 1970)

$$\theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta \tag{9}$$

Although the two assumptions have two different roles they are both needed for identification purposes. The first one (8) is necessary in order to give a scale to the latent variables (Bollen, 1989) and is standard practice in the CFA framework. The second assumption (9) was proposed by Wiley and Wiley (1970) in their seminal paper. The authors suggest that this assumption is sound theoretically as the random error is a product of the measurement instrument and not of the population. And, albeit this assumption has been previously criticised (e.g. Alwin, 2007, p.107) it is still less restrictive than that proposed by Heise (1969), namely that the reliability should be considered equal in time.

Given the previous equations and the definition of reliability in CTT, the percentage of variance explained by the true score (Lord and Novick, 1968), I propose the following measures of reliability for each of the four waves:

$$\kappa_1 = 1 - \frac{\theta}{\psi_{11} + \theta} \tag{10}$$

$$\kappa_2 = 1 - \frac{\theta}{\beta_{21}^2\psi_{11} + \psi_{22} + \theta} \tag{11}$$

$$\kappa_3 = 1 - \frac{\theta}{\beta_{32}^2(\beta_{21}^2\psi_{11} + \psi_{22}) + \psi_{33} + \theta} \tag{12}$$

$$\kappa_4 = 1 - \frac{\theta}{\beta_{43}(\beta_{32}^2(\beta_{21}^2\psi_{11} + \psi_{22}) + \psi_{33}) + \psi_{44} + \theta} \tag{13}$$

where κ_i represents reliability, ψ_{11} is the variance of the true score T_1 and ψ_{22}, ψ_{33} and ψ_{44} are the variances of the disturbance terms. These equations highlight that the total variance at a given time is a combination of random error, time specific true score variance, variance of the true score of the previous waves and stability. These formulas will be used in order to evaluate the impact of the mixed modes on reliability at the different waves.

2.2.2 Latent Markov Chain

Although the QMSM provides a reliability estimate for continuous and ordered variables it cannot do so in the case of discrete, unordered, variables. In this case a more appropriate model would need to take into account each cell of the variables. Such a model was applied to reliability analyses in panel data by Clogg and Manning (1996) and can be considered a Latent Markov Chain model based on the Langeheine and van de Pol (2009) typology. For simplicity I will consider all variables dichotomous although the model can be easily be extended to variables with more categories. I will also assume that the true score has the same number of categories as the observed one, this being a typical approach to these types of models (van de Pol and Langeheine, 1990; Clogg and Manning, 1996; Langeheine and van de Pol, 2009).

Let i, j, k and l be the levels a dichotomous variable A measured at four points in time: A_1, A_2, A_3 and A_4. By levels I refer to the observed response to the item (e.g., answering ’yes’ may be level 1 and ’no’ 2). The cell probability $(ijkl)$ is denoted by $\pi_{A_1A_2A_3A_4}(ijkl)$. The observed tabulation of A_1, A_2, A_3 and A_4 can be explained by a latent variable, X, that has t levels. Thus, $\pi_{A_1A_2A_3A_4X}(ijklt)$ represents the probability of a cell $(ijklt)$ in an indirectly observed contingency table. The model can be written as:

$$\pi_{A_1A_2A_3A_4}(ijkl) = \sum_{t=1}^{T} \pi_{A_1A_2A_3A_4X}(ijklt) \tag{14}$$

The first assumption of such a model is called *local independence* (Lazarsfeld and Henry, 1968). This implies that once we have controlled for the latent variable there is no relationship between the observed variables:

$$\pi_{A_1A_2A_3A_4X}(ijklt) = \pi_X(t)\pi_{A_1|X=t}(i)\pi_{A_2|X=t}(j)\pi_{A_3|X=t}(k)\pi_{A_4|X=t}(l) \tag{15}$$

where $\pi_X(t)$ is the probability that $X = t$, $\pi_{A_1|X=t}(i)$ is the probability $A_1 = i$ conditional on $X = t$ (i.e., $Pr(A = i|X = t)$) and so on.

For the moment the model is analogous to a CFA with one latent and four observed variables. The main difference between the two is that the latent class model does not make any assumption about the distribution of the variables. This model can be extended to a autoregressive one (i.e., quasi-simplex) with four latent variables:

$$\begin{aligned}\pi_{A_1A_2A_3A_4}(ijkl) &= \sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\sum_{t_3=1}^{T}\sum_{t_4=1}^{T} \pi_{X_1}(t_1)\pi_{A_1|X_1=t_1}(i)\pi_{X_2|X_1=(t_1)}(t_2)\pi_{A_2|X_2=t_2}(j)\pi_{X_3|X_2=(t_2)}(t_3)\\ &\quad \pi_{A_3|X_3=t_3}(k)\pi_{X_4|X_3=(t_3)}(t_4)\pi_{A_4|X_4=t_4}(l) \qquad (16)\end{aligned}$$

where $X_1 - X_4$ are the true scores at the four time points, $\pi_{A_i|X_i=t_i}(i)$ is the measurement model (i.e., the relationship between the latent variable and the observed variable at time i) and $\pi_{X_i|X_{i-1}=(t_{i-1})}(t_i)$ is the transition probability from $i - 1$ to i (i.e., stability in time of the true score).

The reliability in this context can be calculated using the conditional odds ratio between X_i and A_i:

$$\Theta_{A_iX_i} = \frac{\pi_{A_i|X_i=1}(1)\pi_{A_i|X_i=2}(2)}{\pi_{A_i|X_i=1}(2)\pi_{A_i|X_i=2}(1)} \qquad (17)$$

where $\Theta_{A_iX_i}$ gives the odds ratio of correct predictions to incorrect ones.

This can be transformed using Yule's Q into a measure of association similar to R^2 (i.e., it is a proportional reduction in error (Clogg and Manning, 1996; Coenders and Saris, 2000; Alwin, 2007)):

$$Q_{A_iX_i} = (\theta_{A_iX_i} - 1)/(\theta_{A_iX_i} + 1) \qquad (18)$$

Thus, $Q_{A_iX_i}$ can be seen as a measure of reliability in the context of LMC as it represents the percentage of the observed variance that is due to the true score as opposed to error.

In order to identify these models two important constraints are needed. The first one is *time-homogeneity of latent transition probabilities* (Alwin, 2007; van de Pol and Langeheine, 1990):

$$\Pi_{X_2X_1} = \Pi_{X_3X_2} = \Pi_{X_4X_3} = \Pi_{X_{t+1}X} \qquad (19)$$

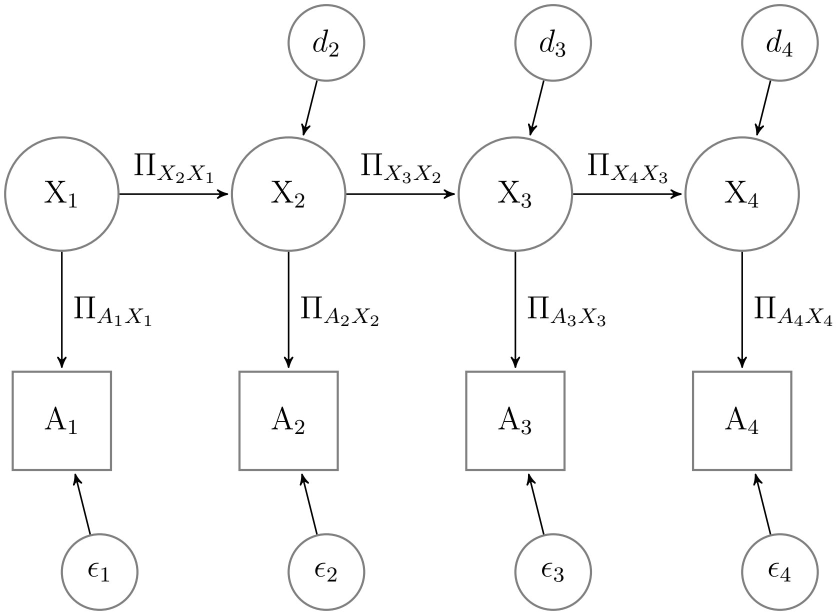
where $\Pi_{X_iX_{i-1}}$ are matrices with transition probabilities of the true scores from one time point to another. The second assumption is that of equal reliabilities over time (Alwin, 2007). Here $\Pi_{A_iX_i}$ are the matrices of conditional probabilities linking the observed and the latent variables:

$$\Pi_{A_1X_1} = \Pi_{A_2X_2} = \Pi_{A_3X_3} = \Pi_{A_4X_4} = \Pi_{AX} \qquad (20)$$

These assumptions imply that, unlike the QMSM, we can only have one estimate of reliability and one of stability[1] for each variable when using LMC. And, even if the two models give similar estimates of reliability, the assumption of equal reliabilities in time of LCM (20) is conceptually different from the assumption of equal error variance in time of the QMSM (9). As a result, the reliabilities of the two types of models will not be compared.

[1] Although equal stability in time may be inappropriate in some situations, e.g., occupation status when the labour market situation changes unexpectedly, this should lead to a similar bias in the two mode designs and should not bias the conclusions.

Figure 2: Latent Markov Chain with four waves



One possible risk of the LMC approach is the resulting high value of the reliabilities. Alwin (2007) highlights that in this kind of model reliability is also a result of the number of categories of the observed variable. Therefore, in the case of items with two categories high levels of reliability are expected. This is not a limitation of the method as long as it can discriminate the mode design effect on reliability and stability.

Concluding the presentation of the two analytical approaches I would also like to highlight that despite the similarity between QMSM and LMC, both conceptually and in one of the assumptions, they are two distinct approaches that come form different statistical traditions (Alwin, 2007). In this paper I see this as an advantage as it gives us two different ways of identifying the impact of mixing modes on measurement.

Furthermore, although I believe that reliability is an important quality indicator it also needs to be highlighted that the models used here ignore the part of the variance that is systematic bias. Although a considerable part of the mixed mode literature talks about types of systematic errors that manifest differently between modes, such as primacy/recency or social desirability (Roberts, 2007; Betts and Lound, 2010; Dex and Gumy, 2011, for an overview), the two models used here, QMSM and LMC, ignore the bias as long as it is stable in time. On the other hand, due to the comparison of the two mode designs part of the systematic and stable error produced by CAPI is controlled for in wave two, because some of the respondents in the mixed mode design responded using CATI. Thus, part of the mode specific systematic bias is transferred to d_2. Keeping in mind this limitation I propose three hypotheses.

2.3 Hypotheses

As motivated in section 2.1 there are three main reasons why mixing modes would lead to a decrease in reliability in the respective wave. Firstly, using a mode that leads to an increase in burden and a decrease in motivation for the respondent will lead to more mistakes and inconsistencies. Furthermore, as long as a mode specific systematic bias exists then the change of mode for a part of the sample will lead to a decrease in reliability by moving this part of variance from the true score into the time specific disturbance term. Lastly, the overall increase in complexity of data collection and processing due to the mixed mode design will lead to the addition of random errors.

H1: The reliability is smaller for the mixed mode design compared to the single mode design in the wave where the former was used.

I also expect a decrease in stability when the mode switches in the mixed mode design. This can be caused by the move of the mode specific variance to either random error or to time specific true score. Thus, for the mixed mode design I expect lower stabilities from wave one to wave two, when some respondents change from CAPI to CATI, and from wave two to wave three, when the same respondents move from CATI to CAPI.

H2: The stability is smaller in waves in which the mode switches, i.e., stability to waves 2 and 3, for the mixed mode design.

Additional impact of mixing modes on reliability is possible in subsequent waves. This effect is important for longitudinal studies as it threatens comparability with previous waves even if the mode switch is temporary. One possible mechanism through which this may take place is panel conditioning. The change of mode leads to a different type of cognitive task which, in turn, stops the increase of reliability of the true scores in subsequent waves.

H3: The reliability will be smaller for the mixed mode design in subsequent waves, even if no design differences remain.

3 Methodology

3.1 Data

The USIP is a yearly panel study that started in 2008 and is financed by the UK Economic and Social Research Council (Understanding Society: Innovation Panel, Waves 1-4, 2008-2011). The survey is used for methodological experiments. It uses a stratified and geographically clustered sample in order to represent England, Scotland and Wales. Using the Postcode Address File it applied systematic random sampling after stratifying for the density of the manual and non-manual population in order to select 120 sectors. Within each of these sectors 23 addresses were selected. The total number of selected addresses was 2.760. In wave 4 a refreshment sample of 960 household was added to the sampling design.

Throughout the survey all residents over 16 were interviewed using Computer Assisted Personal Interviews. In the present analyses I will be using waves 1-4, which have been collected between 2008 and 2011. Wave 1 had an initial household level response rate of 59.5% followed by conditional response rates of 72.7%, 66.7% and 64%, respectively, for subsequent waves (McFall et al., 2012). The individual sample size for the full-interview vary from a maximum of 2384 in wave 1 to a minimum of 1621 in wave 3.

One of the characteristics that were manipulated in the experiments of the USIP is the mode design. For example, in wave two of the survey a CATI-CAPI sequential mixed mode design was implemented for two thirds of the sample and a CAPI single mode design was used for the last third. Furthermore, the sequential design was equally divided in an 'telephone light' group and a 'telephone intensive' group. In the case of the former if one individual from the household refused or was unable/unwilling to participate over the telephone the entire family was transferred to a CAPI interview while in the latter group such a transfer was made only after trying to interview all adults from the household using CATI (Burton et al., 2010). Although this design decision is interesting I will consider the two CATI approaches together and will refer to them as the CATI-CAPI mixed mode design as opposed to the CAPI single mode design.

Because the allocation to the mode design was randomized we can consider the resulting data as having a quasi-experimental design. Using the notation introduced by Campbell and Stanley (1963) I can represent the data as seen in Table 2. The two groups have similar mode design with the exception of wave 2, when the CATI-CAPI sequential design was introduced for a portion of the sample. In addition, the two groups are randomized, as a result they should be comparable and all differences between them should be caused by the mode design.

Table 2: Quasi-experimental design of mixed modes in USIP

Group	W_1	W_2	W_3	W_4
R_{CAPI}	O_1	O_2	O_3	O_4
$R_{CATI-CAPI}$	O_1	XO_2	O_3	O_4

In order to evaluate the impact of the mixed-mode design on the reliability of the items I have selected all the items that were measured in the USIP in all four waves. A Stata .ado file that automatically evaluates the names of the variables in all four waves was used. Additional rules for selecting variables were applied. As a result, all variables that had less than 100 cases for each wave on the pooled data were eliminated. Variables that are not the direct results of data collection (e.g., weighting) or variables without variance (i.e., one category with 100%) were also eliminated.

After this selection and the elimination of nominal variables[2] a total of 46 variables remained. Out of these 18 are analysed using QMSM and 28 dummy variables using LCM.

[2] As reliability and stability are also caused by the number of categories comparisons with the dummy variables would be questionable. And while dichotomizing and analysing these using LMC is an option the process of constructing different categories and comparisons has a high degree of arbitrariness and may not correspond to the substantial uses of the data.

And while the dummy variables cover a wider range of topics, from beliefs and self-description to income and job, the metric and ordinal variables are concentrated on certain themes. The ordinal variables are mainly composed of the SF12, a health scale that measures both physical and psychological well-being (Ware et al., 2007). The continuous variables, on the other hand, measure total income, net and gross, self-description, namely height and weight, and the number of hours worked in a typical week. Each of these 46 variables will be analysed using one of the two methods presented above in order to estimate differences in reliability and stability between the two mode designs.

Table 3: Characteristics of the variables

	Beliefes/ attitudes	**Household**	**Income**	**Job**	**Other**	**Self-description**	**Sum**
Dummy	1	8	2	9	6	2	28
Metric	0	0	2	1	0	2	5
Ordinal	0	0	0	0	1	12	13
Sum	1	8	4	10	7	16	46

The data management and part of the analyses were made using Stata 12. The bulk of the analyses were done using Mplus 7 and the runmplus .ado.

3.2 Analytical approach

For both types of analytical approaches I used BIC to compare the different models:

$$BIC = -2ln(L) + kln(n) \tag{21}$$

where k is the number of free parameters to be used and n is the sample size. This information criterion controls both for sample size and model complexity. Moreover, it does not assume the models are nested and it can be used consistently both for the QMSM and LMC. With this measure a smaller value represents an improvement in model fit as it minimizes the log likelihood.

Before exploring more the ways in which mode influence measurement I need to highlight an important caveat. Although theoretically it makes sense to distinguish between measurement and selection effects in mode differences these are harder to distinguish empirically. A small number of articles have tried to do this so far (Vannieuwenhuyze et al., 2010; Lugtig et al., 2011; Vannieuwenhuyze and Loosveldt, 2012; Buelens et al., 2012). Usually they do so either through a very complex survey design (e.g. Buelens et al., 2012) or by using a number of assumptions (e.g. Lugtig et al., 2011; Vannieuwenhuyze and Loosveldt, 2012). In order to simplify the analyses I will not distinguish between measurement and selection effects. Using the random allocation to mode the total effect of the mixed mode design can be estimated. As a result, differences between the two mode designs in reliability can be seen as a total effect that includes selection, measurement and their interaction.

3.2.1 Quasi-Markov Simplex Model

The QMSM models will be analysed in a sequential order from the most general, less restricted, to the most constrained model. The first model (*Model 1*) assumes that the unstandardised loadings are equal to one (8) and that random measurement error is equal in time (9) within mode design. Thus, nothing is constrained equal across the two mode designs. The next four models stem from the definitions of the reliabilities for the four time points. As a result, *Model 2* assumes that ψ_{11} and θ are equal across modes. If this is true then the reliability for wave one (κ_1) is equal across modes. *Model 3* adds to this constraint the equality of β_{21} and ψ_{22} across modes, implying that κ_1 and κ_2 are equal across modes. The last two models, *Model 4* and *Model 5*, follow a similar logic and constrain β_{32} and ψ_{33}, respectively β_{43} and ψ_{44} to be equal across the two mode designs. Because I expect the biggest differences in wave two, then *Model 3* should not lead to improvement in goodness of fit. If, on the other hand, the best fitting model is *Model 5* then both reliability and stability are equal across modes. Normally, *Model 2* could be used as a randomization test. If the selection of the two groups was indeed random no significant differences for ψ_{11} and θ_1 would be expected across mode designs. Unfortunately, due to the assumption of equal random measurement in time (9), θ is 'contaminated' by the random measurement errors of the rest of the time points. As a result, the model cannot be used as a randomization test.

Although QMSM represents one of the best models we have for measuring reliability with repeated items it is marred with estimation issues. Two of these are the negative variances of some of the variables and standardised stability coefficients over 1.0 (Jagodzinski and Kuhnel, 1987; Van der Veld and Saris, 2003). While Coenders et al. (1999) and Jagodzinski and Kuhnel (1987) explore the causes of these issues I propose a possible solution here. Instead of estimating the models using Maximum Likelihood methods I employ Bayesian estimation. This has the advantage that it needs smaller sample sizes and does not results in unacceptable coefficients (Congdon, 2006). Although these advantages are important the Bayesian estimation has two drawbacks: it cannot use weights and multigroup comparisons have not to been implemented in the software used. The latter is especially important as I aim to compare the two mode designs. In order to bypass this issue I have taken advantage of the fact that this estimation algorithm can deal with missing data using the Full Information procedure (Enders, 2010; Muthén and Muthén, 2012). Using this approach all the information in the data is used for the analysis. We can take advantage of this and model two parallel QMSM for the two groups, although there are no common cases, by imposing the lack of any relationship between them[3]. I will be using the Bayesian implementation in Mplus 7 with the following parameters: four chains, thinning coefficient of five, convergence criteria of 0.01 and a maximum of 70000 iterations and a minimum of 30000 (Muthén and Muthén, 2012).

3.2.2 Latent Markov Chain

The estimation procedure for LMC will include three distinct models. These start once again from the least restrictive and progresses to the most restrictive model. As a result,

[3] Analyses were carried out to compare the Bayesian approach with Maximum Likelihood (with and without weights and a balanced sample). The models resulted in similar estimates of reliability and stability.

Model 1 will assume that both the transition probabilities in time and the reliabilities are equal in time within mode design (19)-(20). *Model 2* imposes the additional restriction that the reliability is the same for the two mode designs (i.e., $\Pi_{AX_{CATI-CAPI}} = \Pi_{AX_{CAPI}}$) and *Model 3* constrains the transition probabilities to be equal across mode designs (i.e., $\Pi_{XX_{t-1_{CATI-CAPI}}} = \Pi_{XX_{t-1_{CAPI}}}$).

By comparing the three models using the BIC we are able to see which model fits the data best. If *Model 1* is the best fitting one then we conclude that both the reliabilities and the transition probabilities from one wave to another (i.e., stabilities) are different across modes. On the other hand, if *Model 3* is the best fitting one we can assume that both the reliability and the stability are equal across the two mode designs. If *Model 2* is the best fitting one we can assume that the reliabilities are equal but the stability of the true scores are not.

In order to estimate the model I will use Robust Maximum Likelihood estimation with 500 maximum number of iterations and random starts: 200 initial stage random starts and 20 final stage optimizations. In order to be consistent I will use no weights but the Full Information procedure will be applied.

4 Analysis and Results

Previous research has highlighted that the QMSM is an unstable model and can sometimes either not converge or give out of bounds coefficients (e.g. Jagodzinski and Kuhnel, 1987; Van der Veld and Saris, 2003). Although using the Bayesian approach bypassed most of these issues it did prove problematic for three of the continuous variables, two items measuring income and one measuring weight. While the models converged when analysed by mode design our parallel quasi-simplex chains approach did not lead to convergence even when increasing the maximum number of iterations or the thinning coefficient. As a result I could compare the reliabilities and stabilities across modes for these variables but I would not be able to use the same approach as presented in section 3.2.1. Consequently, these three variables will be ignored in the following analyses. Similar issues have arisen in the case of LMC. Out of the initial 28 items ten of them have issues in convergence, involving either a non-positive definite first-order derivative product matrix or a non-positive definite Fisher information matrix. One of the solutions proposed, increasing the number of random starts, did not prove successful in any of the models. The items were concentrated on two main topics. Four of them were measuring attributes linked with the household and were derived from household level information. Four of the items were measuring job and income related aspects, such as whether the respondents are full-time or part-time employed. These ten variables will also be ignored in the following analyses. Therefore, our actual variable sample size is 33, 13 being ordinal variables, two continuous and 18 dichotomous.

The sample sizes of the analyses are moderately high because of the Full Information procedure. Thus, for QMSM the median is 1790 and the minimum 1020. On the other hand, the sample sizes are somewhat smaller for the LMC, reaching 534 cases for a variable measuring if the respondent is living in the household with the partner, but still with a median of 1775 individuals included per analysis.

4.1 Quasi-Markov Simplex Model

Concentrating on the 15 ordered variables, 12 of them measure health-related aspects while the other three measure height, number of work hours and when they last weighed themselves. Each of these items was analysed five times, each time imposing a new constrain as presented in section 3.2.1, this procedure resulting in 75 models. Within each variable I compared the BIC of the five models. A decrease of this coefficient indicating an improvement in the model fit while controlling for sample size and model complexity.

Looking at the mean goodness of fit of the models as constrains are added I observe that moving from *Model 1* to *Model 2* leads to a mean decrease in BIC of 33. Similar results are found by adding the constraints of *Model3*. Adding the mode equality of *Model 5* to *Model 4* leads to a further mean BIC improvement of 27. Overall, each constrain leads to improvement of fit and usually *Model 5* proves to be the best fitting one. This implies that there is no difference between the two mode designs in reliability or stability for the ordered variables.

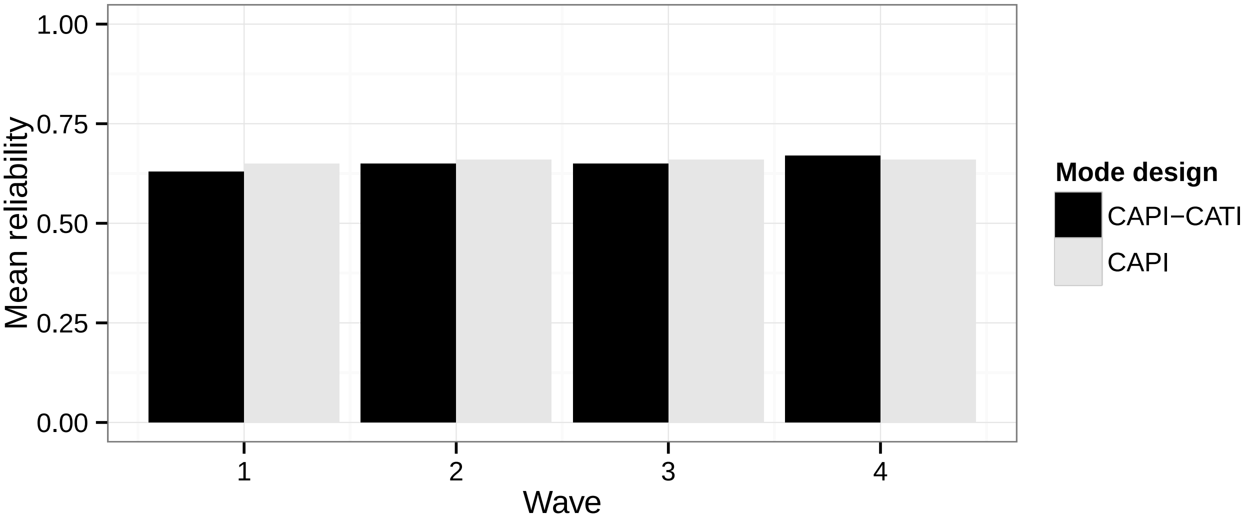
Table 4: BIC differences within variables

Variable	Model	BIC	Difference
Height	Model 1	16328.1	0.0
	Model 2	16323.0	5.1
	Model 3	16337.3	-14.2
	Model 4	16323.3	13.9
	Model 5	16336.3	-13.0
Job hours	Model 1	20655.6	0.0
	Model 2	20647.1	8.4
	Model 3	20664.1	-16.9
	Model 4	20638.8	25.2
	Model 5	20633.4	5.4
SF4b	Model 1	13226.1	0.0
	Model 2	13215.8	10.3
	Model 3	13204.6	11.2
	Model 4	13208.0	-3.3
	Model 5	13195.0	13.0
SF5	Model 1	16473.1	0.0
	Model 2	16473.3	-0.2
	Model 3	16443.6	29.7
	Model 4	16431.7	11.9
	Model 5	16427.9	3.8

Table 4 presents the exceptions to the linear decrease in BIC with the additional constrains. If we look in the sequence of models for the best fitting one and consider that as the best representation of the data then Height is the only variable that does not have *Model 5* as the best fitting model. In this case *Model 2* appears to be the most appropriate representation of the data. Therefore, in the case of Height either the reliability or the stability to wave 2 is different between the two mode designs. Looking in more detail at the estimates of *Model 2* for height we observe that although reliabilities are very similar, 0.974 for the single mode design versus 0.976 for the mixed mode, the difference in the stability of the true score from wave one to wave two is bigger, being 0.966 for the former and 0.997 for the latter. Therefore, it appears that the stability of the Height variable from wave 1 to wave 2 is significantly higher in the CATI-CAPI mixed mode design than in the CAPI design.

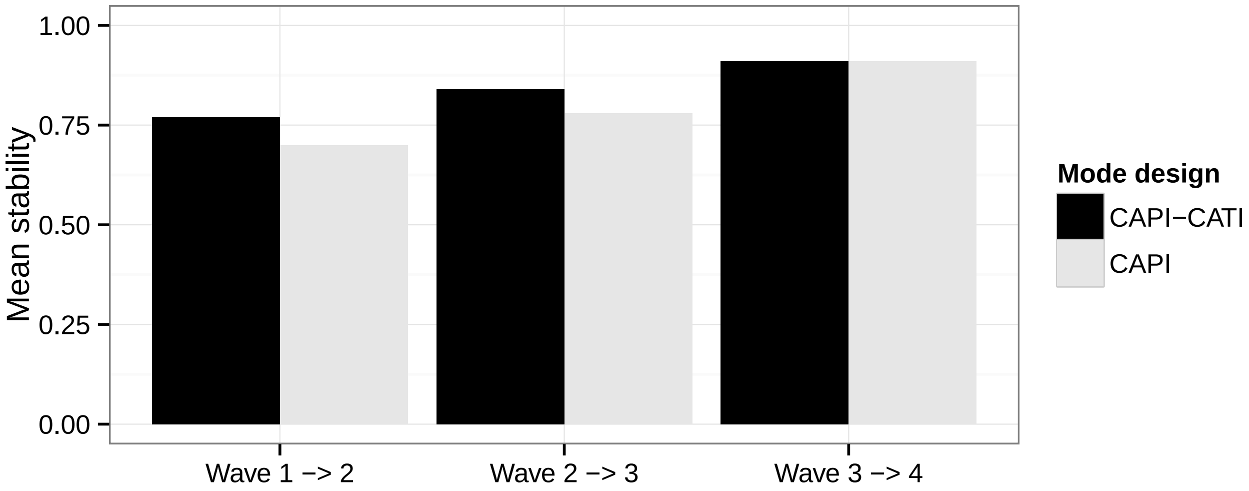
A somewhat similar pattern is indicated by the other three variables presented in Table 4, although they point to *Model 5* as the best fitting model. For example, in the case of *Model 2* for Job hours we see that even if the single mode design shows somewhat larger reliability for wave 2, 0.931 versus 0.924, the stability from wave 1 to wave 2 for the mixed mode design is considerably higher, 0.867 versus 0.726. Similarly, in the case of *Model 3* of SF4b reliability in wave 3 is higher for the CAPI design, 0.566 as opposed to 0.445, but the stability from wave 2 to wave 3 is lower, 0.580 versus 0.940. Similar results can be seen for SF5 for wave 1 in *Model 1*, although with smaller differences.

Figure 3: Mean reliability ordered variables (*Model 1*)



Looking at the overall reliability patterns we observe very small differences between the groups with a moderate mean level of reliability for all the ordered items analysed. Additionally, figure 4 shows the change over time in the mean stability of the items. Here we also find very small differences between the groups, with an overall increase of stability in time. This is an expected result and can be explained both in terms of panel conditioning (Sturgis et al., 2009) and as a selection in time of 'good' respondents (Brehm, 1993). Running the same analyses on a balanced panel led to similar increase in stability over time. This provides an argument for panel conditioning as opposed to selection.

Figure 4: Mean stability ordered variables (*Model 1*)



4.2 Latent Markov Chain

In addition to the QMSM models I have analysed 18 dichotomous variables. For each of these I estimated three models, as presented in Section 3.2.2, resulting in 54 models. Overall, similar results have been found. In mean the constrains of *Model 2*, equal reliabilities in time, brings an average improvement in BIC of 18. A similar result appears when the additional constrain of equal stability across modes designs is imposed. The linear improvement of fit with the two additional constrains is true for all the variables analysed.

Looking at the mean reliabilities and stabilities we find a similar results as in the case of QMSM. The models indicate high reliabilities that are consistent across the two mode designs. For both of them the mean reliability is 0.98. A similar conclusion can be reached in the case of stability. On average the mixed-mode group had a stability of 7.4 while the one for the single mode design was 9.5 on a log odds scale. As the BIC results indicate, this difference does not withstand and the best fitting model is one that constrains them to be equal.

5 Conclusions and discussion

Section 2.1 argued that mixing modes will have a detrimental impact on reliability, especially when one of the modes included brought additional respondent burden and lead to a decrease in motivation. The results of our analyses do not confirm this hypothesis. In the case of QMSM I have found only one variable out of 15 that did not indicate *Model 5* as the best fitting one. A similar result was found when using LMC. Here *Model 3* was always the best fitting one, indicating once again that stability and reliability are equal between mode designs. This implies that for almost all the variables analysed here the reliability and stabilities were equal across modes.

By using the QMSM I was also able to analyse the impact of mixing modes on subsequent waves with regards to reliability. I have argued in section 2.1 that mixing modes may lead to a decrease (or lack of increase) in reliability compared to a single mode design. One potential explanation for such an effect is panel conditioning, the mixing of modes leading to a different type of cognitive task that, in turn, would decrease the impact of training. Our results do not support this hypothesis. No differences in reliabilities between the two mode designs in waves 3 and 4 are observed. The result of no differences across modes regarding panel conditioning is the first one of its kind, to the knowledge of the author, and may indicate that at least on this dimension and for these types of variables, longitudinal reliability, panel studies are 'safe' from mode effects.

Furthermore, the second hypothesis has also been rejected by the data. Through the change of the modes used by some of the respondents an impact on stabilities was expected. The two mode switches implied by the mixed mode design, from CAPI to CATI (weave 1 to wave 2) and from CATI to CAPI (wave 2 to wave 3), did not have a significant impact on the stability of the true score. This can be either due to the lack of differences between the two groups or because the model already takes into account the random error characteristic to each mode design.

Looking in more detail at the panel conditioning I have found mixed results. The finding of constant reliability in time is an unexpected one as previous research has shown effects of panel conditioning (e.g., Jagodzinski et al., 1987; Sturgis et al., 2009). But although an effect of panel conditioning on reliability was not present there was one on stability. Thus, stability of the true scores increases in time even if no mode differences are apparent. Because similar results were found when a balanced panel was analysed conditioning appears more plausible than selection.

Although the overall results in the QMSM indicate that reliability and stability are similar across the two mode designs there are a few exceptions worth mentioning. Firstly, only one variable did not indicate *Model 5* as the best fitting one. In this case the higher stability in the mixed mode design seems to be the main driver. Similarly, three other variables did not show linear improvement of fit although *Model 5* still was the best fitting one. In these cases a pattern of higher reliability for the single mode design versus higher stability for the mixed mode design appeared. This is an unexpected result and further research is needed in order to see if this is a substantially important result or an artefact of the statistical modelling.

Although the results are not definitive and further replications are needed these results indicate that reliability may not be the main threat to cross mode designs comparisons. If these results are replicated then selection (Lynn and Kaminska, 2012; Vannieuwenhuyze and Révilla, in press) and response styles (e.g., Jäckle et al., 2006) may prove to be more important issues than reliability. Although our analyses show that random error is the same in the two mixed mode designs the same cannot be claimed about systematic error that is stable in time (e.g., Billiet and Davidov, 2008). In order to capture this variance alternative approaches are needed, such as MTMM (Saris et al., 2004) or modelling of response styles (Billiet and McClendon, 2000; Billiet and Davidov, 2008).

The study has also contributed to the methodological field by proposing two important solutions to some of the estimation issues that have marred QMSM (Jagodzinski and Kuhnel, 1987; Van der Veld and Saris, 2003). Firstly, I have proposed Bayesian estimation as a way to avoid out of bounds coefficients. This has proved successful as all the models that used this approach converged with coefficients inside the theoretical limits. In addition, a solution to the lack of multi-group modelling when using this estimation method has been proposed. Taking advantage of the Full Information method used for missing data I have modelled two parallel quasi-simplex chains and constrained all covariances between them to zero. This has proved successful for all but three items. Although these have converged when analysed by mode they did not when using this method. More research is needed to understand exactly why this happened.

A series of limitations of the study also need to be highlighted. Firstly, I do not make the distinction between selection and measurement effects but talk about the total effect of mixing modes. Using the random allocation to the design I am able to show the total effects of mixing modes. These results are correct as long as the measurement and selection effects do not impact reliability in opposite directions. Furthermore, I cannot say anything about the decomposition into measurement and selection effects.

Another limitation refers to the modelling approach used here. The QMSM modelling may result in the overestimation of reliability if response styles are stable in time. Previous research has indicated that this may be the true in some cases. For example, Billiet and Davidov (2008) show that the acquiescence factor modelled using two balanced sets of items tapping Distrust in Politics and Perceived Ethnic Threat is stable in time. If this is true for response styles that affect the items tested here then the QMSM model may provide overestimated reliability coefficients. Although this may be an important threat in normal analytical designs it should be highlighted that our conclusions are biased only if the response style stability is different for the two mode designs.

Additionally, out results are also confounded with the different attrition patterns created by the mixed mode design in wave 2. Previous results have shown that the two mode designs lead to different response rates and some minor differences in attrition patterns and response bias (Lynn, 2012). And although the Full Information method assumes Missing At Random this is true only if the missing mechanism is included in the model (Enders, 2010). In the models used here it implies that the missing pattern respects a 1-lag Markov chain. If this is not true and the unexplained missing is linked with reliability then it will confound our results. In order to gauge the degree to which response rates and attrition may be issues I have compared our results to those from using a balanced panel. No differences were apparent.

Another potential limitation of the study may be the high levels of reliability and stability in LMC. These bring doubts regarding its usefulness as an instrument for measuring data quality for dichotomous variables. Even if it is very attractive due to the lack of distributional assumptions it may also prove not sensitive enough to find differences across groups, especially where big discrepancies are not obvious. Nevertheless, the model has previously been able to find heterogeneity between groups (e.g., van de Pol and Langeheine, 1990) and the results found here may only be caused by the small differences across the variables compared (Clogg and Manning, 1996; Langeheine and van de Pol, 2009). This last argument being also supported by the general consistency of the LMC with the QMSM.

Finally, I believe that this type of analysis should be extended to cover more attitudinal and sensitive questions as these may prove to be more susceptible to mode effects. Additionally, analysis of subgroups, such as those with low cognitive abilities or language skills, may lead to higher differences. Similar extensions should also be made in different types mixed-mode designs and different cultural backgrounds.

References

Alwin, D. F. (2007). *The Margins of Error: A Study of Reliability in Survey Measurement.* Wiley-Blackwell.

Betts, P. and Lound, C. (2010). The application of alternative modes of data collection on UK governent social surveys. literature review and consultation with national statistical institutes. *Office for National Statistics*, pages 1–83.

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

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

Bollen, K. (1989). *Structural Equations with Latent Variables.* Wiley-Interscience Publication, New York.

Brehm, J. O. (1993). *The Phantom Respondents: Opinion Surveys and Political Representation.* University of Michigan Press, Ann Arbor.

Buelens, B., van der Laan, J., Schouten, B., van den Brakel, J., Burger, J., and Klausch, T. (2012). Disentangling mode-specific selection and measurement bias in social surveys. *Discussion paper Statistics Netherlands*, pages 1–29.

Burton, J., Laurie, H., and Uhrig, N. (2010). Understanding society innovation panel wave 2 results from methodological experiments. *Understanding Society Working Paper Series*, (04):1–34.

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

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

Chang, L. and Krosnick, J. A. (2009). National surveys via rdd telephone interviewing versus the internet comparing sample representativeness and response quality. *Public Opinion Quarterly*, 73(4):641–678.

Clogg, C. and Manning, W. (1996). Assesing reliability of categorical measurements using latent class models. In Eye, A. v. and Clogg, C., editors, *Categorical Variables in Developmental Research: Methods of Analysis*, pages 169–182. Academic Press Inc.

Coenders, G. and Saris, W. E. (2000). Testing nested additive, multiplicative, and general Multitrait-Multimethod models. *Structural Equation Modeling: A Multidisciplinary Journal*, 7(2):219–250.

Coenders, G., Saris, W. E., Batista-Foguet, J. M., and Andreenkova, A. (1999). Stability of three-wave simplex estimates of reliability. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(2):135–157.

Congdon, P. P. (2006). *Bayesian Statistical Modelling.* Wiley, 2 edition.

Couper, M. (2012). Assesment of innovations in data collection technology for undersanding society. Technical report, Economic and Social Research Council.

De Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, 21(5):233–255.

de Leeuw, E. D. and de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A., editors, *Survey Nonresponse*, pages 41–54. Wiley-Interscience, New York, 1 edition.

Dex, S. and Gumy, J. (2011). On the experience and evidence about mixing modes of data collection in large-scale surveys where the web is used as one of the modes in data collection. *National Center for Research Methods Review Paper*, pages 1–74.

Dillman, D. (2009). Some consequances of survey mode changes in longitudinal surveys. In Lynn, P., editor, *Methodology of Longitudinal Surveys*, pages 127–140. John Wiley & Sons.

Dillman, D. A., Smyth, J. D., and Christian, L. M. (2008). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method.* Wiley, 3 edition.

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

Groves, R. and Kahn, R. (1979). *Surveys by telephone : a national comparison with personal interviews.* Academic Press, New York.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, 75(5):861–871.

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

Holbrook, A. L., Krosnick, J. A., Moore, D., and Tourangeau, R. (2007). Response order effects in dichotomous categorical questions presented orally: The impact of question and respondent attributes. *Public Opinion Quarterly*, 71(3):325–348.

Jäckle, A., Roberts, C., and Lynn, P. (2006). Telephone versus Face-to-Face interviewing: Mode effects on data quality and likely causes. report on phase II of the ESS-Gallup mixed mode methodology project. *ISER Working Paper*, (41):1–88.

Jagodzinski, W. and Kuhnel, S. M. (1987). Estimation of reliability and stability in Single-Indicator Multiple-Wave models. *Sociological Methods & Research*, 15(3):219–258.

Jagodzinski, W., Kuhnel, S. M., and Schmidt, P. (1987). Is there a "Socratic effect" in non-experimental panel studies?: Consistency of an attitude toward guestworkers. *Sociological Methods & Research*, 15(3):259–302.

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

Krosnick, J. A., Narayan, S., and Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New directions for evaluation*, 1996(70):29–44.

Langeheine, R. and van de Pol, F. (2009). Latent markov chains. In Hagenaars, J. and McCutcheon, A., editors, *Applied Latent Class Analysis*, pages 304–341. Cambridge University Press, 1 edition.

Lazarsfeld, P. F. and Henry, N. (1968). *Latent structure analysis.* Houghton, Mifflin.

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Addison-Wesley Publishing Company, Inc.

Lugtig, P. J., Lensvelt-Mulders, G. J. L. M., Frerichs, R., and Greven, F. (2011). Estimating nonresponse bias and mode effects in a mixed mode survey. *International Journal of Market Research*, 53(5):669–686.

Lynn, P. (2012). Mode-switch protocols: how a seemingly small design difference can affect attrition rates and attrition bias. *ISER Working Paper*, (28):1–17.

Lynn, P. and Kaminska, O. (2012). The impact of mobile phones on survey measurement error. *Public Opinion Quarterly.*

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

McFall, S., Burton, J., Jäckle, A., Lynn, P., and Uhrig, N. (2012). Understanding society – the UK household longitudinal study, innovation panel, waves 1-4, user manual. *University of Essex*, pages 1–66.

Muthén, L. and Muthén, B. (2012). *Mplus User's Guide. Seventh Edition.* CA: Muthén & Muthén, Los Angeles.

Révilla, M. (2010). Quality in unimode and Mixed-Mode designs: A Multitrait-Multimethod approach. *Survey Research Methods*, 4(3):151–164.

Révilla, M. (2011). Impact of the mode of data collection on the quality of survey questions depending on respondents' characteristics. *RECSM Working Paper*, (21):1–25.

Roberts, C. (2007). Mixing modes of data collection in surveys: A methodological review. *NCRM Methods Review Papers.*

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

Sturgis, P., Allum, N., and Brunton-Smith, I. (2009). Attitudes over time: The psychology of panel conditioning. In Lynn, P., editor, *Methodology of longitudinal surveys*, pages 113–126. Wiley, Chichester.

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

Understanding Society: Innovation Panel, Waves 1-4, 2008-2011. University of Essex: Institute for Social and Economic Reserch.

van de Pol, F. and Langeheine, R. (1990). Mixed markov latent class models. *Sociological Methodology*, 20:213.

Van der Veld, W. and Saris, W. (2003). A new framework and model for the survey response process. Unifying P. Converse, C. Achen, and J. Zaller, and S. Feldman. pages 1–29, Marburg.

Vannieuwenhuyze, J., Loosveldt, G., and Molenberghs, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, 74(5):1027–1045.

Vannieuwenhuyze, J. T. A. and Loosveldt, G. (2012). Evaluating relative mode effects in Mixed-Mode surveys: Three methods to disentangle selection and measurement effects. *Sociological Methods & Research*, 42(1):82–104.

Vannieuwenhuyze, J. T. A. and Révilla, M. Evalauting relative mode effects on data quality in Mixed-Mode surveys. *Survey Research Methods*.

Ware, J., Kosinski, M., Turner-Bowker, D. M., and Gandek, B. (2007). *User's Manual for the SF-12v2 Health Survey*. QualityMetric, Incorporated.

Wiley, D. and Wiley, J. (1970). The estimation of measurement error in panel data. *American Sociological Review*, 35(1):112–117.

Annex - Variables for which reliability was estimated

ID	Code	Label	Measurement Level	Topic	Analysis
1	**height_i**	Height in inches	Metric	Self-description	Successful
2	**jbhrs**	Thinking about your (main) job, how many hours, excluding overtime and meal breaks, are you expected to work in a normal week?	Metric	Job	Successful
3	**payg_dv**	Gross pay per month in current job: last payment	Metric	Income	No convergence
4	**payn_dv**	Net pay per month in current job: last payment	Metric	Income	No convergence
5	**weight_p**	Weight in pounds	Metric	Self-description	No convergence
6	**SF1**	In general, would you say your health is...	Ordinal	Self-description	Successful
7	**SF2a**	The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much? First, moderate activities, such as moving a table, pushing a vacuum cleaner, or playing golf...	Ordinal	Self-description	Successful
8	**SF2b**	Climbing several flights of stairs...	Ordinal	Self-description	Successful
9	**SF3a**	During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of your physical health?	Ordinal	Self-description	Successful
10	**SF3b**	Were limited in the kind of work or other activities...	Ordinal	Self-description	Successful
11	**SF4a**	During the past 4 weeks, how much of the time have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?	Ordinal	Self-description	Successful
12	**SF4b**	Did work or other activities less carefully than usual...	Ordinal	Self-description	Successful
13	**SF5**	During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)...	Ordinal	Self-description	Successful
14	**SF6a**	These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past 4 weeks… Have you felt calm and peaceful...	Ordinal	Self-description	Successful
15	**SF6b**	Did you have a lot of energy...	Ordinal	Self-description	Successful
16	**SF6c**	Have you felt downhearted and depressed ...	Ordinal	Self-description	Successful

ID	Code	Label	Measurement Level	Topic	Analysis
17	**SF7**	During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting friends, relatives, etc.)...	Ordinal	Self-description	Successful
18	**hlwtl**	When was the last time you were weighed using scales, either by yourself or someone else?	Ordinal	Other	Successful
19	**aidhh**	Is there anyone living with you who is sick, disabled or elderly whom you look after or give special help to (for example, a sick, disabled or elderly relative/husband/wife/friend etc)?	Dichotomous	Household	Successful
20	**aidxhh**	Do you provide some regular service or help for any sick, disabled or elderly person not living with you?	Dichotomous	Household	Successful
21	**caruse**	Do you normally have access to a car or van that you can use whenever you want to?	Dichotomous	Other	Successful
22	**ccare**	Do you ever use any type of childcare for your child/children? By 'childcare' I mean care carried out by anyone other than yourself (or your partner).	Dichotomous	Other	No convergence
23	**cohab_dv**	Lives with cohabitee in hh	Dichotomous	Household	No convergence
24	**drive**	Do you have a full UK driving licence?	Dichotomous	Other	No convergence
25	**employ**	Which description on this card [comes closest to what you first did after leaving full-time education]/[best describes what you did next, even if it was only for a month?]	Dichotomous	Job	Successful
26	**health**	Do you have any long-standing physical or mental impairment, illness or disability? By 'long-standing' I mean anything that has troubled you over a period of at least 12 months or that is likely to trouble you over a period of at least 12 months.	Dichotomous	Self-description	Successful
27	**hlwte**	Are you fairly sure of your weight or is that an estimate?	Dichotomous	Self-description	Successful
28	**j2has**	Do you currently earn any money from a second job, odd jobs, or from work that you might do from time to time, apart from any main job you have?	Dichotomous	Job	Successful
29	**jbft_dv**	Full or part-time employee	Dichotomous	Job	No convergence
30	**jbhas**	Can I just check, did you do any paid work last week - that is in the seven days ending last Sunday - either as an employee or self employed?	Dichotomous	Job	Successful

ID	Code	Label	Measurement Level	Topic	Analysis
31	**jboff**	Even though you weren't working did you have a job that you were away from last week?	Dichotomous	Job	No convergence
32	**jbsemp**	Are you an employee or self-employed?	Dichotomous	Job	No convergence
33	**jbterm1**	Leaving aside your own personal intentions and circumstances, is your job...	Dichotomous	Job	Successful
34	**julk4wk**	Have you looked for any kind of paid work or government training scheme in the last four weeks?	Dichotomous	Job	Successful
35	**julkjb**	Although you are not looking for paid work at the moment, would you like to have a regular paid job even if only for a few hours a week?	Dichotomous	Job	Successful
36	**livesp_dv**	Lives with spouse in hh	Dichotomous	Household	Successful
37	**livewith**	May I just check, are you/is [NAME] living with someone in this household as a couple?	Dichotomous	Household	Successful
38	**lkmove**	If you could choose, would you stay here in your present home or would you prefer to move somewhere else?	Dichotomous	Other	Successful
39	**mobuse**	Do you personally have a mobile phone?	Dichotomous	Other	Successful
40	**opecl30**	Do you believe that people in the UK will be affected by climate change in the next 30 years?	Dichotomous	Beliefes/attitudes	Successful
41	**paytyp_d**	How is your pay calculated, in particular are you salaried or paid by the hour?	Dichotomous	Income	Successful
42	**payusl**	Your take home pay last time was £[PAYNL if PAYNL>0 / PAYGL IF PAYNL=0]. Is this the amount you usually receive (before any statutory sick pay or statutory maternity, paternity or adoption pay)?	Dichotomous	Income	No convergence
43	**respf16_dv**	Whether natural/adoptive/step/foster father of child under 16. Based on edited information collected in the household grid	Dichotomous	Household	No convergence
44	**respm16_dv**	Whether natural/adoptive/step/foster father of child under 16. Based on edited information collected in the household grid	Dichotomous	Household	No convergence
45	**single_dv**	Flag for whether or not respondent lives without a partner in the household	Dichotomous	Household	No convergence
46	**xpmove**	Even though you may not want to move, do you expect you will move in the coming year? / Do you expect you will move in the coming year?	Dichotomous	Other	Successful