Evaluating mode differences in longitudinal data

Moving to a mixed mode paradigm of survey methodology

Alexandru Cernat

A thesis submitted for the degree of

Doctor of Philosophy in Survey Methodology

Institute for Social and Economic Research

University of Essex

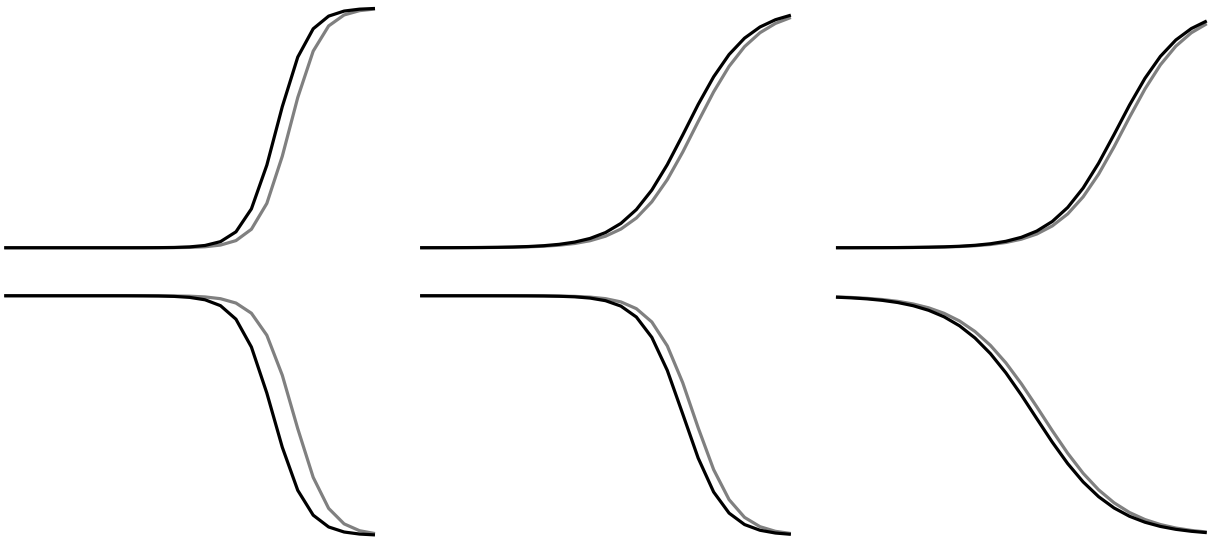
United Kingdom

October 2015

Evaluating mode differences in longitudinal data

Moving to a mixed mode paradigm of survey methodology

Alexandru Cernat



Declarations

No part of thesis has been submitted for another degree.

All work in this thesis is my own.

The chapters in this PhD have been published both as working papers and in peer-review journals. Additionally, two of them have been co-authored. Here I will present the publications and my co-authors contributions to them.

Chapter 2 is sole authored and been published as:

- Cernat, A. (2015). 'Impact of mode design on reliability in longitudinal data'. *Sociological Methods & Research*, 44(3), 427-457.
- Cernat, A. (2013). 'Impact of mode design on reliability in longitudinal data'. *ISER Working Paper*, 2013-09, 1-30.

Chapter 3 is sole authored has been published as:

- Cernat, A. (2015). 'Impact of mixed modes on measurement errors and estimates of change in panel data'. *Survey Research Methods*, 9(2), 83-99.
- Cernat, A. (2014) 'Impact of mixed modes on measurement errors and estimates of change in panel data'. *Understanding Society Working Paper*, 2014-05, 1-21.

Chapter 4 was co-authored with Mick Couper and Mary Beth Ostedal from the Institute for Social Research, University of Michigan and is currently under review at the *Journal for Survey Statistics and Methodology*. I have done the data management and analysis and wrote the first draft of the measurement models, data management and analytical approach, results and conclusions. Mick has wrote the first draft of the introduction, theory and theoretical expectations while Mary Beth wrote the description of the data. Additionally, we have worked together in revising the paper. It has been published as:

- Cernat, A., Couper, M. and Ofstedal, M. B. (2015). 'Estimation of mode effects in the Health and Retirement Study using measurement models'. *ISER Working Paper Series*, 2015-19, 1-18.

Chapter 5 was co-authored with Peter Lynn and is currently under review at the *Social Science Computer Review*. I have done the data management and analysis and wrote the first draft of the data description, results and conclusions. Peter has wrote the first draft of the introduction and theory and has given feed-back on the analysis. Additionally, we have worked together in revising the paper. It has been published as:

- Cernat, A. and Lynn, P. (2014). 'The role of email addresses and email contact in encouraging web response in a mixed mode design'. *Understanding Society Working Paper*, 10, 1-15.

Chapter 6 was sole authored and is under review at *Sociological Methods & Research*. It has been published as:

- Cernat, A. (2015).'Using equivalence testing to disentangle selection and measurement in mixed modes surveys'. *Understanding Society Working Paper*, 2015-01, 1-16.

I dedicate this to my family and my dear ones.

Acknowledgements

I am incredibly grateful for the last three years. It has been a great opportunity to learn and grow, both as a researcher and as an individual. ISER is an outstanding institution that through the great atmosphere of support and understanding enabled me to experiment and develop. I am also grateful to the ESRC that made these three years possible and also enabled me to experience for three months the research and teaching environment in the US, at the University of Michigan.

I still remember receiving three and a half years ago the first mail from Peter Lynn telling me how he would be interested in research on this topic called mixed modes. It was the first time I have heard of the existence of such a research topic. Since then he has guided me from the initial explorations to what I hope to be a contribution to this field. I will always be grateful for his gentle guidance and support. I am also thankful to the guidance offered by Nick Allum, my second supervisor. It was thanks to him that I have ended up at ISER, a path that will influence me for the rest of my life. And while his name is not on the supervisor list I would like to thank Paul Clarke who in the short period he has been here has become a third supervisor, giving invaluable feed-back and supporting me in a future career in methodology.

In the last three years I have started to meet the people behind this blurry research field called methodology. I have come to realize that it is a great company to have: friendly, relaxed, innovative and supportive. I consider myself very lucky to be part of this group. In this period I was also very lucky to start collaborating with a number of these wonderful people. I cannot emphasise enough how important this has been for me to develop and learn. So I thank them here for their willingness to start working with this strange eastern European kid: Mick Couper, Daniel Oberski, Peter Lugtig, Noah Uhrig, Nicole Watson, Mary Beth Ofstedal. Additionally, the conversations with Tarek Al Baghal, Jon Burton, Annette Jackle and Jorre Vannieuwenhuyze have been very helpful for the development of this PhD. Lastly, I also want to thank Gundi Knies, Alita Nandi and Jakob Petersen for the invaluable

comments regarding the Understanding Society data.

And while methodology is an enticing intellectual pursuit these last three years would not have been the same without the great people I have met here and which I like to call my friends: Wouter Zwysen, Simon Cole, Feifei Bu, Natasha Crawford, Carlos Lagorio, Yamil Nares, Caroline Carney, Duygu Ozdemir, Ipek Mumcu, Vilma, Andrea Geraci, Stefanie Hoherz, Elisa Sibley, Angus Holford, Federico Zilio.

None of this would have been possible without the support of my family for which I will always be grateful. And lastly I want to thank Oana for her unwavering faith in me and for her stoic endurance with my existential dilemmas.

Summary

Collecting and combining data using multiple modes of interview (e.g., face-to-face, telephone, Web) is becoming common practice in survey agencies. This is also true for longitudinal studies, a special type of survey that applies questionnaires repeatedly to the same respondents. In this PhD I investigate if and how collecting information using different modes can impact data quality in panel studies.

Chapters 2 and 3 investigate how a sequential telephone - face-to-face mixed mode design can bias reliability, validity and estimates of change compared to a single mode. In order to achieve this goal I have used an experimental design from the Understanding Society Innovation Panel. The analyses have shown that there are only small differences in reliability and validity between the two modes but estimates of change might be overestimated in the mixed modes design.

Chapter 4 investigates the measurement differences between face-to-face, telephone and Web on three scales: depression, physical activity and religiosity. We use a quasi-experimental (cross-over) design in the Health and Retirement Study. The results indicate systematic differences between interviewer modes and Web. We propose social desirability and recency as possible explanations.

In Chapter 5 we investigate using the Understanding Innovation Panel if the extra contact by email leads to increased propensity to participate in a sequential Web - face-to-face design. Using the experimental nature of our data we show that the extra contact by email in the mixed mode survey does not increase participation likelihood.

One of the main difficulties in the research of (mixed) modes designs is separating the effects of selection and measurement of the modes. Chapter 6 tackles this issue by proposing equivalence testing, a statistical approach to control for measurement differences across groups, as a front-door approach to disentangle these two. A simulation study shows that this approach works and highlights the bias when the two main assumptions don't hold.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

These are very interesting times to be a survey methodologist. When this PhD started big data was something that computer scientists did, online opt-in panels was a big business but still not "scientific enough" to threaten the traditional polling establishment and most surveys were still uni-mode. Three years have past and so many things have changed. Now the oldest master program in survey methodology, at the University of Michigan, teaches a course on big data analystics, the University of Essex has started a master program on this topic and the Royal Statistical Society actively seeks to become relevant to data scientists (see meeting organized by the RSS on the 11^{th} of May 2015 for example). Meanwhile polling has moved more and more to online opt-in panels. In arguably the most visible event in the field during the PhD's period the New York Times has moved its polling to YouGov. This has started an open conflict between the "establishment", represented by the American Association of Public Opinion Research (AAPOR) president Michael Link, and the "new wave" represented by Andrew Gelman, one of the most prominent statisticians in the United States. This discussion culminated with AAPOR's stance being compared to that "of Buggy-Whip Manufacturers taking a strong stand against internal combustion engine".

This amazing change of scenery not only begs the question if a PhD started a few years back is still relevant today but also if survey methodology as a field is still significant. Here I will take the stand that was so well articulated in the key note speech of the 5^{th} European Survey Research Association (ESRA) conference by Mick Couper (2013). The first part of the answer is that survey methodology has

developed a number of skill-sets that will be invaluable in the foreseeable future. The ability to ask questions in a way that ensures validity, the ability to think critically of different types of measurement and selection errors, the ability to sample or to use advanced statistically models, all of these will remain valuable. Nevertheless, in order to survive as a field survey methodology must adapt. To do this it must use the best tools available at the moment and be willing to learn new approaches and think outside the box. As it stands, with all the innovations in big/found data or opt-in panels there are still large number of questions that can't be answered and groups of people that can't be reached using the "new wave" of approaches. While this remains true, "traditional" designs are relevant.

One way in which survey methodology has tried to adapt to the changing environment in recent years has been by collecting data using mixed modes approaches. This means combining different modes, such as face-to-face, Web and telephone in order to save costs and compensate each others weaknesses. This has become an essential topic in contemporary methodology. For example, at the ESRA conference in 2013 there were 10 session on mixed modes and 11 related in one way or another with modes (mostly linked to web and mobile surveys). That still holds true at ESRA 2015 (9 and 11 respectively). In the long run this might be just the first design feature that survey methodology may mix. We can easily imagine a future, and this is already happening, in which research questions are answered by mixing big data and surveys, or in which opt-in panels use high-quality probabilistic surveys to correct for their estimates.

It is in this larger context that the work of this PhD sees the light of day. As survey agencies are pressured to save costs and be innovative mixing modes has become a way to do both things in the "traditional" framework of survey methodology. This is also true at the host institution of this PhD, the Institute for Social and Economic Research, University of Essex. While the aim of managing a project such as the UK Household Longitudinal Study (UKHLS, Buck and McFall, 2012), a large representative household panel survey and UK's largest social science research

resource investment, is to have the highest data quality possible it did not make it impervious to such pressures (Couper, 2012). This has translated in a research program that experimented and evaluated different types of mixed modes designs in the Innovation Panel. This is a representative sub-sample of the main-stage, UKHLS survey, used for methodological experiments. Similar initiatives have also be seen in other longitudinal (e.g., British Household Panel Survey or the German Socio-Economic Panel) and cross-sectional surveys (e.g., the European Social Survey).

Why are mixed modes important?

One would wonder why have (mixed) modes designs received so much attention. In the end it is but one aspect in a plethora of design decisions that survey managers have to make on a regular basis. There are a number of reasons why this is the case. The most important one is that it has a major impact on other aspects of data collection, influencing how data collection agencies' structure and how resources should be distributed. In addition to the impact on the survey organization, it also has an important effect on the interaction with respondents and, as a result, on the probability of them participating in the survey and on the quality of their answers (De Leeuw, 2005). Furthermore, decisions regarding modes can have a big impact on costs through the indirect effects on interviewer employment, training, transport, programming, etc. Because of these reasons survey agencies and methodologists want to inform decisions by gathering information on the quality/costs trade-offs in mixed modes designs.

And while the pressures to move to a mixed mode design increases it is becoming ever more relevant for longitudinal surveys as well. When this PhD started only one longitudinal survey, to the knowledge of the author, continuously implemented a mixed mode data collection: the Health and Retirement Study (using a concurrent telephone - face-to-face approach). Meanwhile studies such as the UKHLS, the UK cohort studies and the UK Labour survey are either considering or have already moved to such a design. In this context there is very limited research on this topic and most of these surveys are traversing uncharted territory.

Research questions

It is in order to inform survey methodologists and data users interested in mixed mode designs, especially in longitudinal data, that this PhD was written. To do this I have tackled what I believe to be three essential research questions. As expected, these are not exhaustive and my contribution to them is only a small part of a growing literature.

All these three questions rest on the fundamental expectation that the mode of interview can bias the way people respond in surveys. This has been found in the previous literature and is based on theoretical expectations regarding how people answer questions (see for an overview: De Leeuw, 2005; Roberts, 2007; Betts and Lound, 2010; Dex and Gumy, 2011; Tourangeau et al., 2000). The first measurement difference across modes that has been found systematically in the previous literature is social desirability bias. It has been shown that some modes facilitate the honesty of respondents by increasing the privacy of the interview and by reducing the disclosure to interviewers (for reviews see Groves et al., 2008; Tourangeau et al., 2000). Thus, respondents tend to have lower levels of social desirability bias in self-administered modes, such as mail, Computer Assisted Self-Interview or Web surveys. Similarly, this bias tends to be higher for interviewer modes, such as telephone or face-to-face. There is limited evidence that telephone surveys can also have higher levels of social desirability bias compared to face-to-face surveys (see Holbrook et al., 2003). The second main cause of differences found between modes has been primacy/recency bias. This refers to the fact that some people may choose the first category (i.e., primacy) regardless of the content of the question when these are presented visually, such as in mail, Web or face-to-face with showcard surveys, while they also tend to select the last category (i.e., recency) in aural modes (see Krosnick and Alwin, 1987; Schwarz et al., 1992; Visser et al., 2000). Lastly, differences in other aspects, such as the degree of motivation, cognitive burden and explanations/clarifications available might also cause differences in measurement across modes. From these theoretical expectations regarding measurement differences across modes stem the three

research questions that, I hope, will be informative both for survey methodologists and practitioners.

Firstly, the PhD aims to **estimate mode (design) effects in the context of panel data**. This means showing how panel data is distinct from cross-sectional surveys and how these unique characteristics might interact with mixed modes designs. Chapters 2 - 4 have tackled this issue by investigating different aspects of the problem. For example, in Chapter 2 I have shown that reliability (i.e., consistency of answers) is the same for 32 out of 33 questions for the first four waves of a face-to-face single mode design and an alternative approach that includes a telephone - face-to-face sequential data collection in one of the waves. Chapter 3 expands on this and investigates alternative indicators of data quality in the same design by analysing the SF12 scale, a health measure. Results indicate that only one of the twelve items shows systematic differences between the designs but also highlights the potential of long-term effects of the mixed mode design in longitudinal studies. Furthermore, the chapter investigates how an essential coefficient in longitudinal data, the estimate of change, could be biased in the mixed mode design. The results indicate that four out of the 12 items overestimate change compared to the single mode approach.

Chapter 4 uses a different strategy to estimate mode effects. Using a (cross-over) quasi-experimental design in the Health and Retirement Study it is possible to compare directly the measurement quality of three scales, depression, physical activity and religiosity between face-to-face, telephone and Internet. Results indicate that the biggest differences can be found between the interviewer modes (face-to-face and telephone) and the Web. Possible explanations such as social desirability and recency are put forward.

The second question tackled in the PhD is how to **improve the design of the mixed mode approach**. This is done indirectly, by evaluating mode differences, which in turn indicate what and how to improve in the design stage in order to minimize these effects. For example, our the findings in Chapter 4 lead us to rec-

ommend survey managers to combine more similar modes, such as telephone and face-to-face without show-cards, or self-completion modes. But the PhD contributed to this question more directly as well. In Chapter 5, for example, we have looked if previously collected information, namely email addresses, can be used to increase propensity to participate in a sequential Web - face-to-face survey. Results indicate that this is not the case but there are indications that it might increase propensity to respond by Web as opposed to face-to-face, thus saving costs.

Finally, the PhD has tried to support the current research in mixed mode data by **proposing a new way to separate selection and measurement mode effects**, one of the main difficulties in this area of research. This has been done in Chapter 6 where I propose conceptualizing equivalence testing, a statistical approach to comparing and correcting for measurement differences across groups, as a way to separate selection and measurement mode effects. I also present the main two assumptions of the model, exhaustiveness and isolation, and how these bias results when they do not hold in the data.

Chapter 2

The impact of mixing modes on reliability in longitudinal studies

Abstract

Mixed mode designs are increasingly important in surveys and large longitudinal studies are progressively moving to or considering such an approach. 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.

2.1 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, 2013) 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 used for 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 to 2/3 of the sample while the rest took part in a CAPI single mode design. This context gives 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 (CTT Lord and Novick, 1968).

2.2 Background

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, non-response 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; De Leeuw, 2005; Couper, 2011). 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 the analysis usually 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 between mode designs may be due to the different proportions 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 in estimates 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.

The reliability of the data in longitudinal studies can be influenced by four distinct factors. The first one 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 (e.g., Holbrook et al., 2003), 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 through 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 auditory 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 increased reliability and stability of items and decrease in item non-response (e.g., Jagodzinski and Kuhnel, 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

Table 2.1: Mixed modes effects on reliability in a panel study

Cause	Mechanism	Waves affected
Simple mode effect	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

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 2.1 summarizes the possible effects of mixing modes on reliability in panel data compared to a single mode design.

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, 2012; Vannieuwenhuyze and Révilla, 2013). 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 Multitrait-Multimethod 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, 2013). 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.

Reliability in panel data

In order to evaluate the effect of the mixed mode design on the data quality I will concentrate on estimating the impact on reliability. Using Classical Test Theory

(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 (CFA Bollen, 1989) 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 have been 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 con-

ceptually 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.

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:

$$A_1 = \lambda_{11} T_1 + \epsilon_1 \tag{2.1}$$

$$A_2 = \lambda_{22} T_2 + \epsilon_2 \tag{2.2}$$

$$A_3 = \lambda_{33} T_3 + \epsilon_3 \tag{2.3}$$

$$A_4 = \lambda_{44} T_4 + \epsilon_4 \tag{2.4}$$

In addition to the measurement part, the model includes a structural dimension which estimates the relationships between the true scores. As a result of the auto-regressive (simplex) change in time of the true scores we have the following equations:

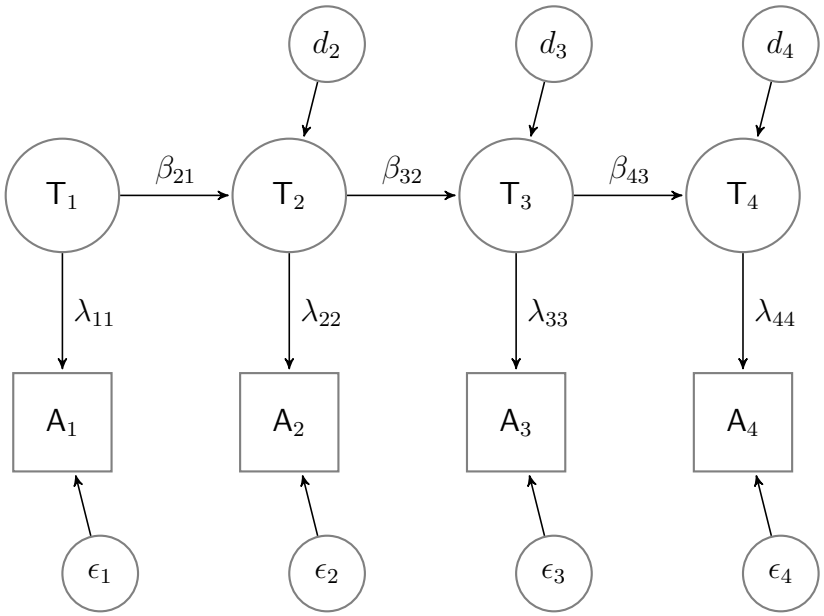
$$T_2 = \beta_{21} T_1 + d_2 \tag{2.5}$$

$$T_3 = \beta_{32} T_2 + d_3 \tag{2.6}$$

$$T_4 = \beta_{43} T_3 + d_4 \tag{2.7}$$

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

Figure 2.1: Quasi–Markov Simplex Model for four waves



model can be seen in Figure 2.1.

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

$$\lambda_{11} = \lambda_{22} = \lambda_{33} = \lambda_{44} = 1 \tag{2.8}$$

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

$$\theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta \tag{2.9}$$

Although the two assumptions have two different roles they are both needed for identification purposes. The first one (2.8) 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 (2.9) 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[1]:

$$\kappa_1 = 1 - \frac{\theta}{\psi_{11} + \theta} \tag{2.10}$$

$$\kappa_2 = 1 - \frac{\theta}{\beta_{21}^2 \psi_{11} + \psi_{22} + \theta} \tag{2.11}$$

$$\kappa_3 = 1 - \frac{\theta}{\beta_{32}^2(\beta_{21}^2 \psi_{11} + \psi_{22}) + \psi_{33} + \theta} \tag{2.12}$$

$$\kappa_4 = 1 - \frac{\theta}{\beta_{43}^2(\beta_{32}^2(\beta_{21}^2 \psi_{11} + \psi_{22}) + \psi_{33}) + \psi_{44} + \theta} \tag{2.13}$$

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.

The QMSM model has a series of assumptions that are needed in order to converge and give correct estimates of reliability and stability. In addition to those mentioned earlier, some of these include: the random errors and the time specific trues scores are not serially correlated, the random errors are not correlated with the true scores, no correlation between the true scores and the random errors, the true scores have a lag-1 time dependence.

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 variable. Such a model was applied to reliability analyses in panel data by

[1]These formulas are equivalent to those put forth by Wiley and Wiley (1970) but are adapted to the model based hypothesis testing that will be presented in Section 2.3.

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 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 of 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, in our case 16, levels. Thus, $\pi_{A_1A_2A_3A_4X}(ijklt)$ represents the probability of a cell $(ijklt)$ in an indirectly observed contingency table. Furthermore, $\pi_X(t)$ can be written to represent the probability that $X = t$ while $\pi_{A_1|X=t}(i)$ is the probability $A_1 = i$ conditional on $X = t$ (i.e., $Pr(A = i|X = t)$), which can also be extended to the other observed variables.

This notation can be included in an autoregressive model (i.e., quasi-simplex) with four latent variables:

$$\begin{aligned}\pi_{A_1A_2A_3A_4}(ijkl) \;=\; & \sum_{t_1=1}^{T}\sum_{t_2=1}^{T}\sum_{t_3=1}^{T}\sum_{t_4=1}^{T}\pi_{X_1}(t_1)\pi_{A_1|X_1=t_1}(i)\pi_{X_2|X_1=(t_1)}(t_2)\pi_{A_2|X_2=t_2}(j) \\ & \pi_{X_3|X_2=(t_2)}(t_3)\pi_{A_3|X_3=t_3}(k)\pi_{X_4|X_3=(t_3)}(t_4)\pi_{A_4|X_4=t_4}(l) \qquad (2.14)\end{aligned}$$

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:

$$\Theta_{A_iX_i} = \frac{\pi_{A_i|X_i=1}(1)\pi_{A_i|X_i=2}(2)}{\pi_{A_i|X_i=1}(2)\pi_{A_i|X_i=2}(1)} \qquad (2.15)$$

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)):

$$Q_{A_iX_i} = (\theta_{A_iX_i} - 1)/(\theta_{A_iX_i} + 1) \tag{2.16}$$

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):

$$\Pi_{X_2X_1} = \Pi_{X_3X_2} = \Pi_{X_4X_3} = \Pi_{X_{t+1}X} \tag{2.17}$$

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:

$$\Pi_{A_1X_1} = \Pi_{A_2X_2} = \Pi_{A_3X_3} = \Pi_{A_4X_4} = \Pi_{AX} \tag{2.18}$$

These assumptions imply that, unlike the QMSM, we can only have one estimate of reliability and one of stability[2] 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 (2.18) is conceptually different from the assumption of equal error variance in time of the QMSM (2.9). As a result, the reliabilities of the two types

[2]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.

of models will not be compared.

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.

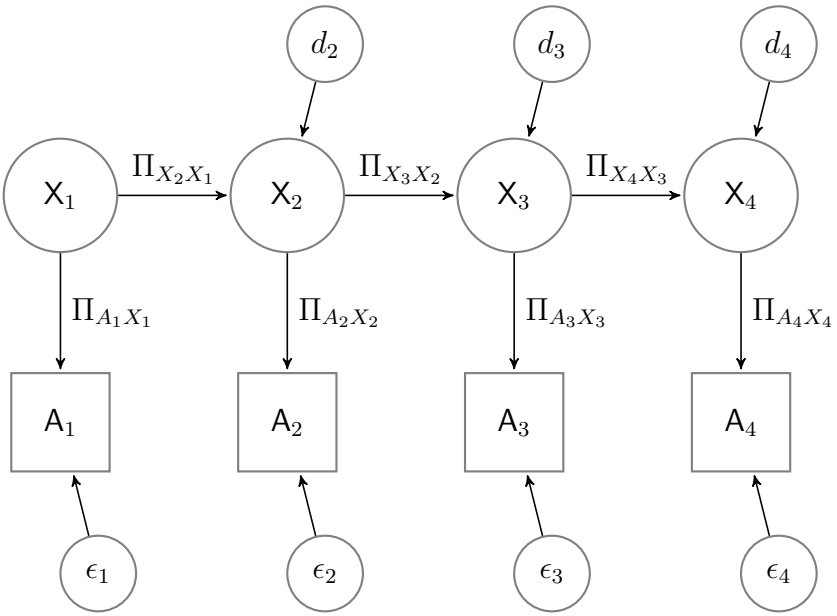
Concluding the presentation of the two analytical approaches I would also like to highlight that despite the similarity between QMSM and LMC, both conceptually and in one of the assumptions, they are two distinct approaches that come form different statistical traditions (Alwin, 2007). In this paper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

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. Thus, part of the mode specific systematic bias is transferred to d_2. Keeping in mind this limitation I propose three hypotheses.

Hypotheses

As motivated in section 2.2 there are four 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

Figure 2.2: Latent Markov Chain with four waves



to a decrease in reliability by moving this part of variance from the true score into the time specific disturbance term. Thirdly, changing modes can have an impact on panel conditioning, thus decreasing reliability and stability. 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 lower 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 lower in the waves in which the mode switches, i.e., stability to waves two and three, 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 can lead to a different type of cognitive task which, in turn, may stop the increase of reliability in subsequent waves.

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

2.3 Methodology

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 occupations 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, consisting of an additional 8 addresses in each of the 120 sectors. Throughout the survey all residents over 16 were interviewed using Computer Assisted Personal Interviews. In the present analysis 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 household response rates conditional on previous wave participation (plus non-contacts and soft refusals in the previous wave) of 72.7%, 66.7% and 69.9%, respectively, for subsequent waves

Table 2.2: Quasi-experimental design of mixed modes in USIP

Group	**Wave 1**	**Wave 2**	**Wave 3**	**Wave 4**
R_{CAPI}	O_1	O_2	O_3	O_4
$R_{CATI-CAPI}$	O_1	XO_2	O_3	O_4

(McFall et al., 2013). The household response rate for the wave 4 refreshment sample was 54.8% (McFall et al., 2013). 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 was 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 a 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 2.2. The two groups have similar mode design (i.e., observations and are noted as O in the table) with the exception of wave 2, when the CATI-CAPI sequential design was introduced for a portion of the sample (highlighted by X in the table). In addition, the two groups are randomized (highlighted in the table by the use of R in the first column), as a result they should be comparable and all differences between them should be caused by the mode design.

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

Table 2.3: Characteristics of the variables

	Beliefs/ 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

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[3] a total of 46 variables remained. Out of these 18 are analysed using QMSM and 28 dummy variables using LCM. 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[4].

The data management and part of the analyses were made using Stata 12. The

[3]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.

[4]All the items analysed here have identical formulation in all the waves. Furthermore, most of them are part of the the core questionnaire and, as such, the respective sections have not changed in time. But, although this is true, some of the other sections and variables in the questionnaire changed across waves. Some of these changes may precede the variables analysed here. This may prove problematic if it has a influence on the random errors and stabilities of the items and of these effects are different across mode designs.

bulk of the analyses were done using Mplus 7 and the runmplus.ado.

Analytical approach

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

$$BIC = -2ln(L) + kln(n) \tag{2.19}$$

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 et al., 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 et al., 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.

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 (2.8) and that random measurement error is equal in time (2.9) 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 the variance of the true score in wave one (ψ_{11}) and the variance of the random error (θ) are equal across designs. If this is true then the reliability for wave one (κ_1) is equal across modes. *Model 3* also constrains the stability of the true score from wave one to wave two (β_{21}) and the variance of the time specific true score in wave two (ψ_{22}) equal across mode designs, implying that the reliabilities of wave one and two (κ_1 and κ_2) are equal across designs. The last two models follow a similar logic. *Model 4* constraints the stability from wave two to wave three (β_{32}) and the variance of the time specific true score of wave three (ψ_{33}). *Model 5* constraints the stability from wave three to wave four (β_{43}) and the variance of the time specific true score in wave four (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 designs. Normally, *Model 2* could be used as a randomization test. If the selection of the two groups was indeed random no significant differences for the variance of the true score (ψ_{11}) and the variance of the random error (θ_1) would be expected across mode designs. Unfortunately, due to the assumption of equal random measurement in time (2.9), the random error (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 and standardised stability coefficients over 1.0 (Jagodzinski et al., 1987; Van der Veld and Saris, 2003). While Coenders et al. (1999) and Jagodzinski et al. (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 yet 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, 2012a). 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[5]. 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, 2012a).

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, *Model 1* will assume that both the transition probabilities in time and the reliabilities are equal in time within mode design (2.17)-(2.18). *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

[5]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.

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.

2.4 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 et al., 1987; Van der Veld and Saris, 2003). Although using the Bayesian approach bypassed most of these issues[6] 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 2.3. 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,

[6]In the case of the ordered variables most of the analyses were done both with Maximum Likelihood estimation and with the Bayesian approach. The former method has proved problematic for almost half of the models. Most of the issues were due to Heywood cases (i.e., negative variances). Usually the variance of the random error was close to 0 and in some cases it ended up being negative. The Bayesian approach has bypassed most of these issues while resulting in similar estimates as the ML estimation. Thus, the Bayesian analysis seems to be a more appropriate approach for the current paper.

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.

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 constraint, as presented in section 2.3. This procedure results 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 constraints 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 *Model 3*. Adding the mode equality of *Model 5* to *Model 4* leads to a further mean BIC improvement of 27. Overall, each constraint 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 2.4 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

Table 2.4: BIC differences within variables

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

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[7] 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 2.4, 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

[7]The stability will be presented as the total variance explained by the previous wave which is equal to $\beta^2_{i,i-1}$.

Figure 2.3: Mean reliability ordered variables (*Model 1*)

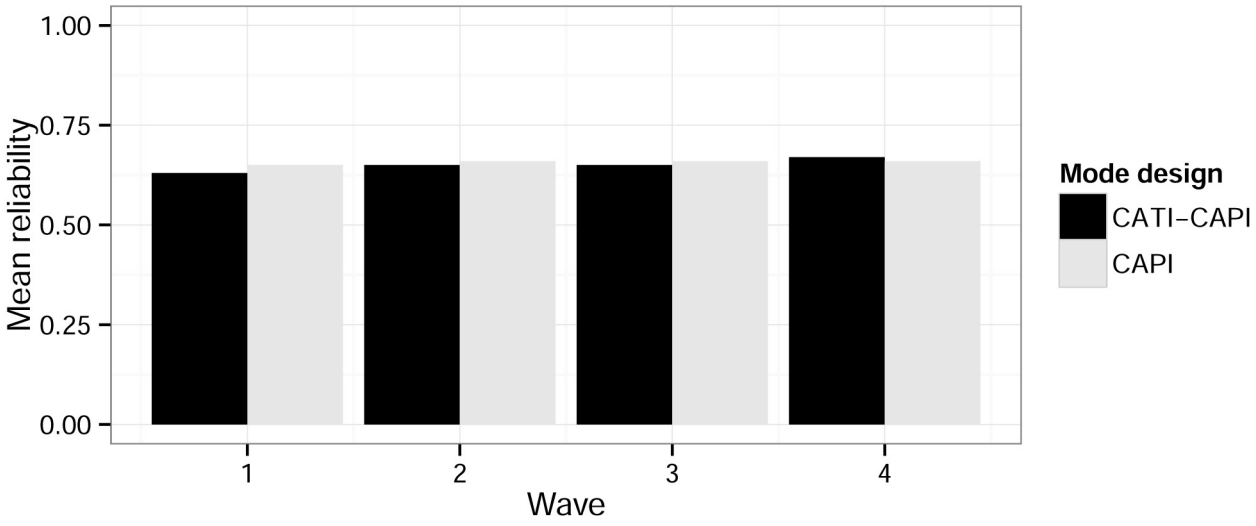
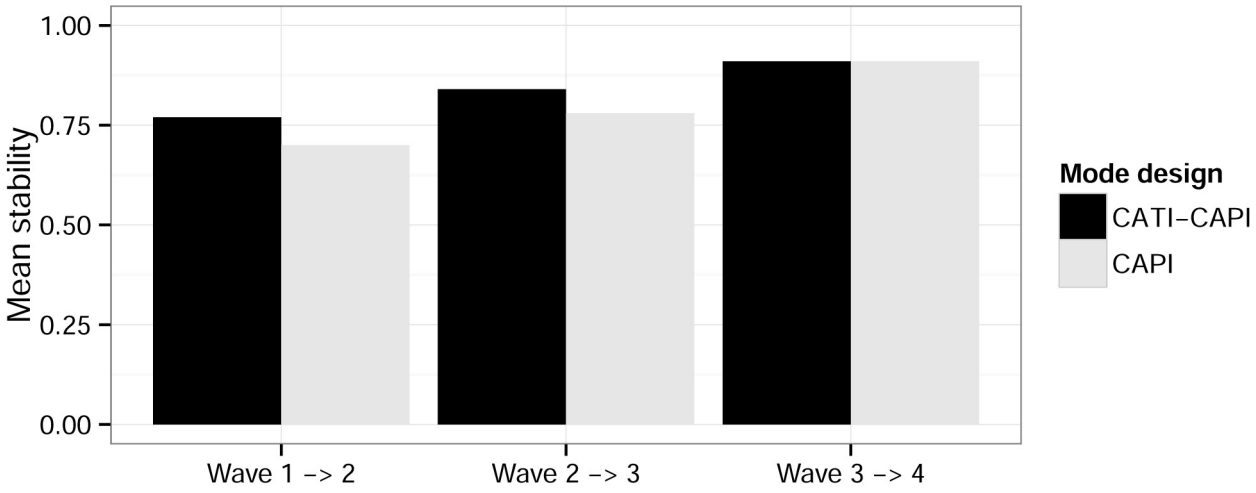


Figure 2.4: Mean stability ordered variables (*Model 1*)



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.

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 2.4 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.

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 2.3, resulting in 54 models. Overall, similar results have been found. On average the constraints of *Model 2*, equal reliabilities in time, brings a mean 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 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. These high levels of stability indicate that there is little time specific change in true score for the variables measured here. This may be caused by a number of factors, two of the most important ones being the fact that change is dependent on the number of categories of the variables (i.e., fewer categories imply smaller probability of change) and that the variables analyzed here may have small degrees of change in time. As the BIC results indicate, the differences between the two mode designs in stability and reliability do not withstand.

2.5 Conclusions and discussion

In Section 2.2 I have argued that mixing modes will have a detrimental impact on reliability, especially when one of the modes 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 2.2 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 mode designs 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, longitudinal reliability, and for these types of variables panel studies are 'safe' from mixed-mode specific effects.

Furthermore, the second hypothesis has also been rejected by the data. A decrease in stabilities was expected because some of the respondents changed the modes used. The two mode switches implied by the mixed mode design, from CAPI to CATI (wave 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 and Kuhnel, 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, 2013; Vannieuwenhuyze and Révilla, 2013, in press) and response styles (e.g., Jäckle et al., 2006) may prove to be more important issues than reliability. Although the 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 Multitrait-Multimethod (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 et al., 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 by using the random allocation to the design. 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, 2013). 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). The models used here imply that the missing pattern respects a 1-lag Markov chain. If this is not true and the unexplained missing data 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, the analyses presented in this paper did not take into account the different subgroups that may be more susceptible to these design changes. As such, possible extensions of this paper can look in more detail at special subgroups, such as respondents with low cognitive abilities or language skills, or at more attitudinal and sensitive questions as these may prove to be more susceptible to mode design effects. Such development should also be encouraged for different types of mixed-mode designs and for different cultural backgrounds.

Chapter 3

Impact of mode design on measurement errors and estimates of individual change

Abstract

Mixed mode designs are receiving increased interest as a possible solution for saving costs in panel surveys, although the lasting effects on data quality are unknown. To better understand the effects of mixed mode designs on panel data we will examine its impact on random and systematic error and on estimates of change. The SF12, a health scale, in the Understanding Society Innovation Panel is used for the analysis. Results indicate that only one variable out of 12 has systematic differences due to the mixed mode design. Also, four of the 12 items overestimate variance of change in time in the mixed mode design. We conclude that using a mixed mode approach leads to minor measurement differences but it can result in the overestimation of individual change compared to a single mode design.

3.1 Introduction

Continuing decreases in response rates, economic pressure and technological advances have motivated survey methodologists to find new solutions for non-response and saving costs. Combining multiple modes of interviews (e.g., telephone, face-to-face, web) has been proposed as a possible solution. This design strategy has also been considered in longitudinal surveys. In the UK, for example, the National Child Development Study 2013 has used a Web Self-Administered Questionnaire–Computer Assisted Telephone Interview (CATI) sequential design while Understand-

ing Society (Couper, 2012) and the Labour Force Survey (Merad, 2012) are planning a move to a mixed mode design. Although these are exciting opportunities for innovation in survey methodology they also provide a number of unique challenges.

Some of these refer to the need for research regarding the effects of mixed modes on selection, measurement and statistical estimates. This is even more urgent in the case of longitudinal surveys as they face specific challenges such as attrition, panel conditioning or estimating change. In the absence of research regarding the potential interactions of these characteristics with mixed mode designs it is not possible to make informed decisions about combining modes in longitudinal surveys. For example, applying a mixed mode design may increase attrition which, in turn, may lead to loss of power and, potentially, higher non-response bias (e.g., Lynn, 2013). Similarly, changing the mode design may bias comparisons in time or estimates of individual change. If such effects are present in the data, the potential benefits of saving costs may be eclipsed by the decrease in data quality.

In order to tackle these issues we will firstly analyze the effect of using a mixed mode design on random and systematic errors in a panel study. This will be done in the wave in which the mixed mode design is implemented and in subsequent waves in order to estimate both the direct and the lasting effects due to mode design. Secondly, we will show how mixing modes influences estimates of individual change in time. The analysis will be based on the first four waves of the Understanding Society Innovation Panel (UKHLS-IP). These data were initially collected using Computer Assisted Personal Interview (CAPI) but they also included a CATI-CAPI sequential design (De Leeuw, 2005) for a random part of the sample in wave two (McFall et al., 2013). The Short Form 12-item Survey (SF12) health scale (Ware et al., 2007) will be used to evaluate the mode design effects.

Previous research on mixed mode designs has concentrated on two main approaches: one that compares *modes* (e.g., CATI versus CAPI) and one that compares *mode design (systems)* (e.g., CATI-CAPI versus CAPI, Biemer, 2001). In the present paper we will use the latter method by taking advantage of the random-

ization into mode design in the UKHLS-IP. Thus, the results will compare mixed modes (sequential CATI–CAPI) to a CAPI single mode design, showing *mode design effects*, as opposed to researching *mode effects*, which would be based on a comparisons of CATI and CAPI that confound measurement and selection.

The paper will present next the main theoretical debates and current research about the two modes included in the design, CAPI and CATI, and mixes of the two. Then, the data, the UKHLS-IP, and the analysis procedure, equivalence testing in Structural Equation Modeling, will be presented. The paper will end with a presentation of the results and a discussion of their implications.

3.2 Background

In order to tackle the issues described above we will first present the theoretical framework and current empirical findings in the literature. Thus, we will highlight differences both between the two modes used, CATI and CAPI, and the impact of mixing these modes on survey results. The last subsection will discuss the specific challenges faced by longitudinal studies and how they can interact with mixed mode designs.

There is a vast literature that compares CAPI and CATI which focuses on two main aspects: selection (i.e., coverage and non-response) and measurement effects (see Groves, 1979, 1990; Groves et al., 1988; Schwarz et al., 1991, for an overview). Due to the data collection design used here we will ignore the debate regarding coverage differences. Using multiple modes in longitudinal studies means that the sampling frame is less problematic as it is possible to use the contact information available in other waves or modes. Thus, this section will concentrate on non-response and measurement differences.

One of the main discrepancies that exist between the two modes used here is the channel of communication: auditory, for CATI, as opposed to both auditory and visual, for CAPI (Krosnick and Alwin, 1987; Tourangeau et al., 2000). These attributes can cause systematic bias such as recency and primacy. While the first

one refers to the favoring of response options that are present at the end of a list and is characteristic for auditory only modes, such as CATI, the latter refers to the preference for the first categories in a list and is a characteristic of visual modes, such as self-completion and interviewer modes with showcards, such as CAPI (Schwarz et al., 1991). A number of studies have shown recency effects in telephone studies (e.g., McClendon, 1991; Holbrook et al., 2007; Bishop and Smith, 2001) while others showed mixed findings regarding primacy effects in self administered modes and face-to-face surveys with showcards (e.g, Bishop and Smith, 2001; Sudman et al., 1996).

An important aspect that differentiates the two modes is the perceived legitimacy of the survey (Tourangeau et al., 2000). This may have an impact both on nonresponse, people having a lower propensity to respond when legitimacy is low, and measurement, causing higher social desirability. Here CAPI has a slight advantage through the use of picture identification badges, written literature and oral presentations given by the interviewer (Groves, 1990). On the measurement part, it is unclear which mode leads to bigger social desirability bias. While CAPI has a slight advantage in legitimacy, disclosure to the interviewer may be easier on the phone due to higher social distance. Previous research on the topic of these modes and social desirability has been mixed (Hochstim, 1967; Groves, 1979; Aquilino, 1992, 1998; Greenfield et al., 2000; Holbrook et al., 2003; Jäckle et al., 2010)

Additionally, satisficing (Krosnick, 1991), the tendency not to engage in thorough cognitive processing of the questions and answers from the survey, may also be different between the two modes. This has two main causes: cognitive burden and motivation. CATI is, on average, conducted at a faster pace (Groves, 1979; Schwarz et al., 1991; Holbrook et al., 2003), thus increasing the burden on the respondent. Also, the absence of visual cues, like showcards or body language, translates into an increased burden compared to CAPI. Furthermore, the motivation can be lower in CATI (Holbrook et al., 2003) as social distance is larger and break-offs are easier. These three phenomena lead to a larger satisficing in CATI

compared to CAPI. This effect can be observed in more random errors, straightlining, 'Don't Know's', acquiescence and other mental shortcuts (Krosnick, 1991) and has been found in previous research focused on comparing the two modes (e.g., Holbrook et al., 2003; Krosnick et al., 1996).

Looking at the overall differences between the two modes, face-to-face and telephone, some consistent results have been found. Face-to-face surveys tend to have slightly higher response rates and smaller non-response bias when compared to telephone surveys (Groves, 1979; Weeks et al., 1983; Aquilino, 1992; Biemer, 2001; Groves et al., 1988; Voogt and Saris, 2005). When analyzing effects on measurement most studies find small or no differences at all (Groves et al., 1988; Greenfield et al., 2000; Aquilino, 1998), with some exceptions (e.g., Biemer, 2001; Jäckle et al., 2010).

These theoretical and empirical differences between face-to-face and telephone modes can become manifest when mixed mode designs are applied. Nevertheless, the way the modes are combined, as well as the decision of modes to be used, can make potential biases harder to predict and quantify. Thus, literature comparing mode designs has found inconclusive results. For example, Link and Mokdad (2006) have shown that combining CATI with web or mail can lead to higher response rates compared to a single mode CATI design. Similarly, Voogt and Saris (2005) have found that combining multiple modes of interview leads to an increase in response rates. These results have not been always replicated. Martin and Lynn (2011a) have shown by using data from an European Social Survey experiment in the Netherlands that a single mode CAPI design achieved a 52% response rates as opposed to 45% for a sequential mixed mode design and 46% for a concurrent one. Also, Olson et al. (2012) have found no differences between single mode mail or CATI designs compared to mail and web mixed mode approach. Looking at non-response bias Klausch et al. (2015) have found that while a CAPI followup can decrease selection bias in some situations, such as in the case of a CATI or a mail survey, it may be less effective in others, such as in the case of a web sample.

Focusing on measurement differences in the context of mixed mode surveys

Révilla and Saris (2010) shows that for some scales, such as social trust, there is no difference between single and mixed modes approaches while for others, such as media and political trust, there are. The results are furthermore complicated in the case of the satisfaction dimension that shows differences both between the two types of data collections and between the two types of mixed mode designs, concurrent and sequential. Nevertheless the differences are not as large as expected, being smaller than the differences between the methods used (Révilla and Saris, 2010). Similarly, Klausch et al. (2013) have found significant differences in data quality between self-administered and interviewer modes but not between CAPI and CATI within a mixed mode survey.

Recent years have seen a development of mixed mode designs and studies to gauge their impact. Starting from the previous literature that compared different modes there are two main approaches. Firstly, part of the literature concentrated on the overall effect of mixing modes on data quality (e.g., Révilla and Saris, 2010; Cernat, 2015b; Lynn, 2013). A separate branch of research has strove to separate measurement and selection effects (e.g., Biemer, 2001; Jäckle et al., 2010; Lugtig et al., 2011; Vannieuwenhuyze et al., 2012; Schouten et al., 2013), most of the time using statistical models to find causal mode effects (e.g., Lugtig et al., 2011; Vannieuwenhuyze et al., 2012).

Mixing modes in longitudinal studies

As mentioned in the introduction, longitudinal studies are different from other surveys in a number of ways. Three main characteristics stand out: attrition, panel conditioning and estimates of individual change. These may, in turn, interact with the mixed mode design. Currently there is very limited research regarding these possible interaction effects.

The first specific challenge when collecting repeated measures from the same individuals is attrition. While this can be considered a specific type of non-response error, it has a number of unique characteristics: it is based on a more stable rela-

tionship between survey organization/interviewer and respondent, and there is the possibility of using previous wave information both for adapting data collection, and for non-response adjustment. The differences between cross-sectional (or first wave) non-response and attrition appear in previous research in this area (Sturgis et al., 2009; Lugtig et al., 2014). This phenomenon can be complicated when combined with a mixed mode design. For example, Lynn (2013) has found that two different mixed mode designs using a CATI-CAPI sequential approach led to different attrition patterns, both compared to each-other and to a CAPI single mode design.

A second issue specific to longitudinal studies is panel conditioning. This process takes place when learning or training effects appear due to the repeated exposure of the respondents to a set a questions/topics. This, in turn, results in an increase over time in the reliability and consistency of responses (Dillman, 2009). Applying mixed mode designs in panel surveys makes this measurement effect unpredictable, as it may interact with the new mode or the way in which the modes are mixed. Presently there is only limited information on how panel conditioning may interact with the mixed mode design. Cernat (2015b) has showed that switching from a CAPI design to a CATI-CAPI sequential approach does not change patterns of reliability and stability, indicating that panel conditioning may not interact with a mixed mode design. Nevertheless, more research is needed to see if this is true using different approaches for measuring conditioning in longer panel studies and for different combinations of modes.

Lastly, panel surveys are especially developed to estimate individual changes in time for the variables of interest. Previous literature has showed that change coefficients are less reliable than the variables that compose them (Plewis, 1985; Kessler and Greenberg, 1981). Their estimation is even more complicated in the case of longitudinal studies that either use a mixed mode design from the beginning or change to such a design in time. Any differences confounded with the new mode(s) or the mixed mode design will bias estimates of change in unknown ways. So far there is no research on this topic.

3.3 Data and methodology

In order to investigate the impact of mixing modes on data quality and estimates of change in panel data we will be using the Understanding Society Innovation Panel. The data is representative of the UK population (England, Scotland and Wales) over 15 and the sampling frame is the Postcode Address File. Here only the first four waves of data (collected one year apart starting from 2008) will be used. The conditional household response rates were 59% (1,489 households), 72.7% (1,122 households), 66.7% (1,027 households) and 69.9% (916 households), respectively, for each of the four waves. The conditional individual response rates were: 84%, 84%, 79% and 79.4%. The fourth wave added a refreshment sample of 960 addresses by applying the same sampling approach. The household response rate for this sample was 54.8% (465 households) while the individual response rate was 80.1% (for more details: McFall et al., 2013) .

The UKHLS-IP was developed to explore methodological questions based on experiments. One of these randomized 2/3 of the sample to a CATI-CAPI sequential design, while the other 1/3 participated in a CAPI single mode design in the second wave. For the rest of the four waves all respondents participated using a CAPI single mode design. Approximately 68% of the respondents in the mixed mode design responded by telephone, while the rest did so using the face-to-face (McFall et al., 2013). Overall, the response rates for the mixed mode design were significantly lower than in the single mode design: 73.9 vs. 65.6 (N=2,555) in wave 2, 65.2 vs. 59.8 (N=2,521) in wave 3 and 57.1 vs. 54.0 (N=2,506) in wave four (for more details: Lynn, 2013).

The UKHLS-IP included a large number of topics, from household characteristics to income sources and health ratings. In order to evaluate the impact of the mixed mode design on measurement errors and estimates of change the SF12 will be analyzed. This scale is the short version of the SF36 and has a wide range of applications, both in health research, and in the social sciences (Ware et al., 2007). The questions and the dimensions/subdimensions that they represented are summarised

Table 3.1: The SF12 scale measures physical and mental health and is based on eight initial subdimensions measured in SF32.

Dimension	Subdimension	Code	Abbreviated content
Physical	General health	SF1	Health in general
	Physical functioning	SF2a	Moderate activity
		SF2b	Climbing several flights
	Role physical	SF3a	Accomplished less
		SF3b	Limited in kind
	Bodily pain	SF5	Pain impact
Mental	Role emotional	SF4a	Accomplished less
		SF4b	Did work less carefully
	Mental health	SF6a	Felt calm and peaceful
		SF6c	Felt downhearted and depressed
	Vitality	SF6b	Lot of energy
	Social functioning	SF7	Social impact II

in Table 3.1. For exact wording and response categories refer to the Annex.

In addition to the fact that the SF12 is widely used and, thus, research based on it would prove useful in a range of fields, analyzing it has some extra advantages. Firstly, it is a scale that is backed up by theory and has been widely tested before. As a result, using it will highlight how mode design differences impact both reliability and validity. Additionally, the scale measures a relatively intimate topic, which may lead to increases in social desirability. This may give us insight in the ways in which the different mode designs may influence aspects such as legitimacy, social distance and trust. Lastly, the scale has both positively and negatively worded questions, which would make differences in acquiescence (i.e., the tendency of selecting the positive answer) more obvious (Billiet and McClendon, 2000).

Equivalence testing

The previous section has revealed that the main focus of mixed modes research is to find causal effects of mode or mode design systems. This can be done either with specific statistical models or with (quasi-)experimental designs. The present paper applies the latter approach in order to measure causal effects of mode design. Due to randomization to mode design we are able to compare the single mode design to the mixed mode design without having to use statistical models for selection. The

remaining task is to compare the two groups. In order to do this we will utilize Structural Equation Modeling (SEM, Bollen, 1989). In this framework, statistically testing differences in coefficients across groups is called equivalence testing.

This approach can be used to compare measurement models across groups. The Classical Test Theory put forward by Lord and Novick (1968) decomposes the observed items into true scores and random errors. Further development has added to this model systematic errors such as method effects (Campbell and Fiske, 1959; Saris et al., 2004; Saris and Gallhofer, 2007b), social desirability (Holtgraves, 2004; Tourangeau et al., 2000) or acquiescence (Billiet and Davidov, 2008; Billiet and McClendon, 2000). Using multiple measures of the same dimension (Alwin, 2007), it is possible to estimate the theoretical concept using a latent variable with Confirmatory Factor Analysis (CFA). In this framework the loading (or slopes) linking the latent variable and the observed variable is the reliability, while the intercepts are the systematic part (Van de Vijver, 2003).

This model can be further extended to categorical observed variables. In such a model a continuous, latent response variable is assumed to exist which determines the observed categories in the data. The answer categories are determined by the relationship between the continuous latent variable and a set of threshold parameters, the number of these coefficient being one less than the number of response categories (for further elaboration see Millsap, 2012).

This model can be incorporated in a Multi Group Confirmatory Factor Analysis when comparing more groups using equivalence (Steenkamp and Baumgartner, 1998; Millsap, 2012; van de Schoot et al., 2012; Meredith, 1993; Byrne et al., 1989). Previous research using this approach has focused on three types of equivalence that can be further extended. The first type is called configural equivalence. If this type of equivalence is found in the data, the structure of the measurement model (i.e., the relationships between latent variables and observed scores) is similar across groups. This can be made more restrictive by assuming metric equivalence, thus implying that the loadings are equal between the groups analyzed. Theoretically, this means

that part of the reliability/random error is the same. Furthermore, the model can also assume that the intercepts are equal across groups, leading to scalar equivalence. This step implies that part of the systematic error is the same across groups. Only when this last type of equivalence is found can the means of the latent variables be meaningfully compared. These three types of equivalence can be extended by constraining more parts of the measurement model to be equal. These can be: the variances of random error, the variances of substantial latent variable, correlations between latent variables or the means of the substantive latent variable.

The procedures used in equivalence testing of multiple groups can also be applied in the case of ordinal variables. Here, the thresholds will be constrained equal across groups in order to test for scalar equivalence, instead of intercepts. In order to estimate the models a number of additional restrictions have to be added to the model. These are presented in the next section. A similar procedure has already been presented and applied in the context of mixed mode research by Klausch et al. (2013).

The measurement model can also be conceptualized as one composed of three parts: random error, systematic error and the substantive part. Thus, differences between groups in loading or variance of random error indicate that there is unequal reliability across groups (Bollen, 1989), the intercept or thresholds are linked to systematic error (Chen, 2008), while the rest of the constraints are linked to substantive variance. Applying equivalence testing to the mode design comparison can make possible the identification of mode design effects on the two types of measurement error. This would help pinpoint the differences between the two designs and indicate possible causes. Furthermore, when the comparison of the groups is supported by randomization, all the differences can be associated with the mode design system (Biemer, 2001).

With SEM it is also possible to estimate individual change in time by using Latent Growth Models (LGM, Bollen and Curran, 2005). These have been developed to estimate both within and between variation and are equivalent to a multilevel

model with a random intercept and slope. The LGM estimates the means for the intercept and slope latent variables (i.e., intercept and a slope for time in a multilevel/hierarchical model), their variances (i.e., random intercepts and slopes for time) and the correlation between the two. Combining the LGM with equivalence testing makes it possible to evaluate the degree to which the estimates of change in time are equal between the groups. When applying this approach to a mode design comparison in panel data, we are able to investigate how much the switch in data collection approach biases individual estimates of change.

Analytical approach

The analysis will be carried out in three main steps. The first one will evaluate, using CFA, the fit of the theoretical model of the SF12 to the UKHLS-IP data. The best-fitting model will be used for the equivalence testing in the second step. This will be done in order to gauge mode design effects in the random and systematic parts of the model. The procedure will be repeated in each of the four waves. The analysis in the first wave will provide a test of randomization, as no differences are expected before the treatment. On the other hand, the equivalence testing in waves three and four will evaluate the lasting effects of mixing modes on the measurement model. Any differences in these waves can be linked to effects of mode design on attrition or panel conditioning. The last stage of the analysis will evaluate the impact of the mixed mode design on estimates of change by testing the equivalence of the LGM for each variable of the SF12.

In order to evaluate the similarity of the SF12 measurement model across mode designs, seven models for each wave will be tested. The cumulative equality constraints applied to the model are:

- *Model 1:* same structure (configural invariance);
- *Model 2:* loadings (metric invariance);
- *Model 3:* thresholds (scalar invariance);
- *Model 4:* error variances (equal random error);

- *Model 5:* latent variable variances;
- *Model 6:* correlations;
- *Model 7:* latent variable means.

The models represent different degrees of equivalence and, as a result, of different mode design effects. Thus, if the best fitting model is *Model 1*, then all the coefficients are different across mode designs. While, at the other extreme, if *Model 7* is the best one, then there are no mode design effects. *Model 4* is an intermediate step and if it is found to be the best fitting one it means that random and systematic error are the same across mode designs, but the substantive coefficients are not.

In order to evaluate the impact of mode design on estimates of change, the third step in the analysis, the following models will be applied to each of the SF12 variables. The cumulative equality constraints applied to the LGM in the two mode designs are:

- *Model 1:* no constraints;
- *Model 2:* slope means;
- *Model 3:* slope variance;
- *Model 4:* correlation between intercept and slope.

Here, again, if *Model 1* is the best fitting model then all the change estimates are different across mode designs, while if *Model 4* is chosen then there are no mode design effects in estimates of change.

The mean and variance of the intercept latent variable will not be tested. Firstly, the mean of the intercept latent variable is assumed to be 0 in the LGM. Secondly, we do not expect any differences at the starting point between the two groups because the same mode design was applied, and selection in the mixed mode experiment was randomized. On the other hand, the equality of the relationship between change in time and the starting point can be tested using *Model 4*.

In order to estimate these models we will be using Mplus 7 (Muthén and Muthén,

2012b) with Weighted Least Squares Means and Variance (WLSMV, Millsap and Yun-Tein, 2004; Muthén et al., 1997; Asparouhov and Muthén, 2010). This estimation approach can take into account the ordinal character of the data. No weighting will be used [1].

Equivalence testing can be complicated when applied to ordinal data. This is true for the variables that are analyzed here. In this case a number of restrictions have to be used. Here we will use the Theta approach (Muthén and Asparouhov, 2002; Millsap and Yun-Tein, 2004). This implies adding the following constraints to the models in order to have convergence:

- all intercepts are fixed to 0;
- each item will have one threshold equal across groups;
- one item for each latent variable will have two equal thresholds across groups;
- for LGM, all the thresholds of the observed items are equal across groups.

For more details about the statistical procedures used for equivalence testing see Millsap and Yun-Tein (2004), Millsap (2012) and Muthén and Asparouhov (2002).

3.4 Analysis and results

The first step of the analysis will explore to what degree the theoretical model of the SF12 is found in the UKHLS-IP. Although the SF12 is widely used both in health and the social sciences, CFA is rarely used to evaluate it. The theoretical model will be tested using the first wave, with the entire sample of UKHLS-IP. Additional relationships, such as correlated errors or cross-loadings, will be added using Modification Indices and goodness of fit evaluation. The final model selected in the first wave will be tested in the next three waves in order to have a confirmatory testing approach and avoid capitalization on chance.

The SF12 theoretical model put forward by Ware et al. (2007) is presented in

[1] The current study is concerned with the overall effect of using a mixed mode as opposed to a single mode design. As such it is focused on how the two samples compare to each other without any other correction. Additionally, the development and use of weights varies considerable by country, data collection agency and field of research. As such, we believe that this approach will provide more generalizable findings.

Figure 3.1. As opposed to the SF36, the subdimensions are only measured by one or two variables (see Table 3.1) and, thus, are not reliable enough to be estimated individually. As a result, the two main dimensions, physical and mental health, will be estimated using latent variables, each with six indicators.

This is the first model tested and presented in Table 3.2 [2]. The model has a moderate fit, with the CFI indicating good fit (Hu and Bentler, 1999), 0.977, while the RMSEA indicating poor fit, 0.103. Using the biggest Modification Indices, which are also theoretically consistent, we add cross-loadings and correlated errors. To ensure that there is appropriate power to identify misspecifications we also calculate the power estimates put forward by Saris et al. (2009) as implemented in the JRule program. The $\Delta\chi^2$ method, difference in χ^2 and degrees of freedom between nested models, is used to test whether the newly added coefficient significantly improves the model. The Mplus function DIFFTEST is used here and the next sections for the $\Delta\chi^2$ method, because of the WLMSV estimation. This uses a model specific correction in the estimation of the $\Delta\chi^2$. For more details refer to: http://www.statmodel.com/chidiff.shtml.

Using this procedure on the first model we identify a cross-loading for SF6b ('Lot of energy') with "Physical" as having the highest Modification Index, 418.9, with an expected value for the parameter 0.76. The power of this test is of 0.768. Freeing this parameter improves the model significantly, leading to a χ^2 of 1143.047 and $\Delta\chi^2$ of 158.828 with 1 degree of freedom. This procedure is repeated until the final model (which is also presented in Figure 3.1) is found. All the new relationships lead to significant improvements in fit and appropriate power is present for all the Modification Indices estimated, these ranging from approximately 0.8 to 1. The final model has a good fit both for RMSEA (0.033) and CFI (0.998) and also fits

[2]The use of fit indicators in the SEM is part of a lively debate that has developed an array of new indicators as well as refute most of them. For example, χ^2 is limited by susceptibility to sample size and deviations from multivariate normality while other indicators, such as RMSEA, have low performances in models with few degrees of freedom (Kenny et al., 2014). In this paper we aim to ameliorate the situation by using a number of fit indicators together as well as evaluating relative improvement in fit as opposed to absolute fit. Thus, the focus here will lies in differences in χ^2 between models as well as improvements in the other fit indicators, namely CFI and RMSEA.

Figure 3.1: The theoretical model of the SF12 does not fit the UKHLS-IP data. A number of cross-loadings and correlated errors are evident in the data.

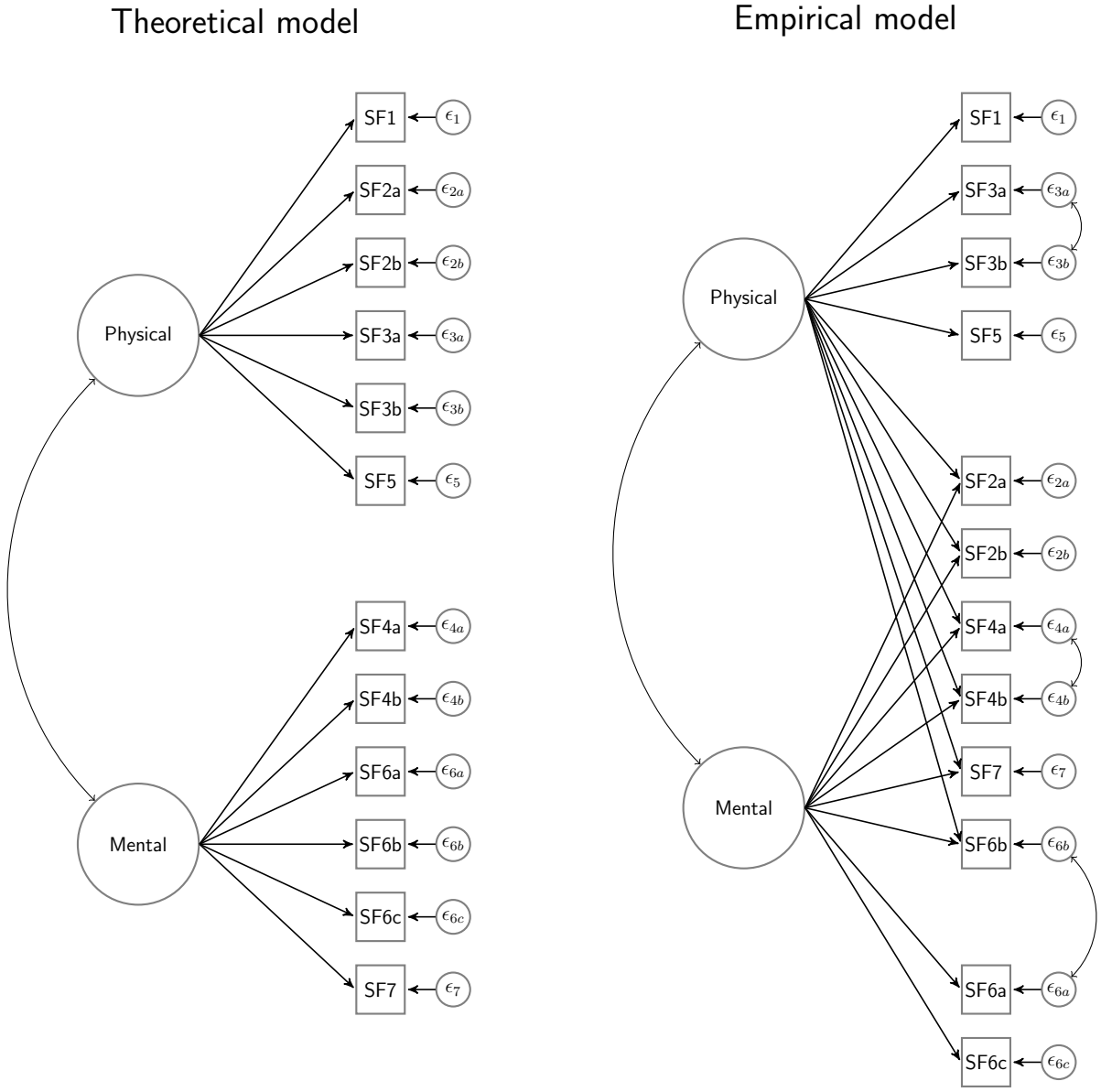


Table 3.2: Model fit after cumulatively adding cross-loadings and correlated errors to the SF12 in wave one of the UKHLS-IP. Final model is also tested in the subsequent three waves.

								Misspecification			
Model	χ^2	**df**	**RMSEA**	**CFI**	$\Delta\chi^2$	**Δdf**	**p**	**Coefficient**	**MI***	**EPC****	**Power**	**NCP*****
Ware et al. 2007	1493.632	53	0.103	0.977				SF6b	418.9	0.76	0.768	7.253
SF6b	1143.047	52	0.09	0.983	158.828	1	0.00	SF7	439.8	-0.735	0.814	8.142
SF7	746.905	51	0.073	0.989	149.789	1	0.00	SF3b with SF3a	178	0.118	1	127.84
SF3b with SF3a	592.933	50	0.065	0.992	86.131	1	0.00	SF4b	151.6	-0.407	0.857	9.152
SF4b	474.691	49	0.058	0.993	59.878	1	0.00	SF6a with SF6b	84.66	0.126	1	53.326
SF6a with SF6b	390.55	48	0.053	0.995	85.316	1	0.00	SF4a	35.02	-0.213	0.793	7.718
SF4a	372.407	47	0.052	0.995	15.855	1	0.00	SF4b with SF4a	148	0.245	0.999	24.66
SF4b with SF4a	230.308	46	0.04	0.997	85.756	1	0.00	SF2b	37.04	-0.147	0.933	11.957
SF2b	199.137	45	0.037	0.998	22.727	1	0.00	SF2a	34.24	-0.129	0.995	20.577
SF2a	170.244	44	0.033	0.998	18.664	1	0.00					
Wave 2	134.751	44	0.033	0.998								
Wave 3	159.45	44	0.043	0.998								
Wave 4	214.178	44	0.045	0.997								

** Modification Indice ** Expected Parameter Change *** Non-Centrality Parameter.*

well in waves two, three and four.

While a number of new relationships have been added to the initial model, most of them have theoretical foundations or have been found in previous research. For example, two of the correlated errors are present between items that measure the same subdimensions: role physical and role emotional. The third correlation, between SF6a and SF6b, has not been found previously but may be due to the similar wording (as in the case of Maurischat et al., 2008) or the proximity. Also, some of the cross-loadings found here were highlighted by previous research on the scale (Resnick and Nahm, 2001; Salyers et al., 2000; Cernin et al., 2010; Rohani et al., 2010). Finally, some of the cross-loadings may be due to the vague words used in the items, which may be associated both with physical and mental health, such as those found in role emotional, vitality and social functioning dimensions.

Equivalence testing across the four waves

Using the model chosen in the previous subsection (empirical model in Figure 3.1) we will test the cumulative constraints of the measurement model across the two mode designs using the sequence presented in Section 3.3. The first wave will be analyzed in order to test the randomization into the treatment. Because everything is the same between the groups in wave one, before the mixed mode design was implemented, no differences are expected in the measurement model. Table 3.3 shows the results of this analysis. The baseline model, which does not impose any equality constraints between the two groups but assumes that the model found in the previous section holds for both, has a good fit with a χ^2 of 189.71, RMSEA of 0.036 and CFI of 0.997. Imposing Metric invariance, equal loadings between groups, does not significantly worsen the model ($\Delta\chi^2$ of 20.3 with 16 df). Repeating the procedure indicates that all constraints hold in wave one of the data, meaning that the measurement model is completely equivalent between the two mode designs. This implies that random and systematic error, but also substantial coefficients like the mean of the latent variables, are equal across the two groups.

Next, the wave two data is analyzed. This is the wave in which the mixed mode design was implemented and where the biggest differences are expected. The results show that the metric equivalence, equal loadings, is reached. The model has a RMSEA of 0.028 and a CFI of 0.998 and a $\Delta\chi^2$ of 20.6 with 13 df. On the other hand, scalar equivalence, equal thresholds, is not reached as the $\Delta\chi^2$ is significant (16.1 with 30 df). By investigating the Modification Indices and the differences in thresholds, SF6a, 'Felt calm and peaceful', is identified as the potential cause. When this threshold is freed the $\Delta\chi^2$ test is not significant (40 with 27 df), indicating that there is partial scalar invariance for all variables except SF6a (Byrne et al., 1989). The rest of the constraints imposed hold, indicating that the only difference in the measurement model between the two mode designs is in the thresholds of SF6a.

Using the same procedure in wave three indicates that metric invariance holds as it does not significantly worsen the Baseline model ($\Delta\chi^2$ 19.7 and 16 df). On the other hand Scalar invariance, equal thresholds, does not hold ($\Delta\chi^2$ 45.7 with 30 df). Investigating the Modification Indices identifies SF4b, 'Did work less carefully', as the potential cause. When all the thresholds are constrained to be equal across groups except SF4b the $\Delta\chi^2$ is not significant anymore (40 with 27 df). Once again, the rest of the coefficients are equal across the two groups. Because the same data collection was used in this wave (i.e., CAPI), differences can only be caused by the interaction of mode design and attrition or panel conditioning.

The evaluation of the fourth wave indicates that there is complete equivalence across the two mode designs. This means that any differences caused by the mode design on the measurement model disappeared after two waves.

Having a closer look at the two significant differences found in the previous analyses reveals that the thresholds for SF6a in wave two are larger for the mixed mode design (Table 3.4). As mentioned before these are indicators of systematic differences between the two designs and are the equivalent of intercepts in continuous Multi Group Confirmatory Factor Analysis. In the categorical analysis the thresholds are indicators of the relationship between the continuous unobserved variable

Table 3.3: The equivalence of the SF12 health scale across mode designs in the four waves of UKHLS-IP is tested. The mixed mode design has an effect on the threshold of SF6a in wave two and in the next wave on SF4b.

Wave	Model	χ^2	df	RMSEA	CFI	$\Delta\chi^2$	df	p
Wave 1	Baseline by mode design	189.71	90	0.036	0.997			
	Metric invariance	185.57	106	0.03	0.998	20.3	16	0.21
	Scalar invariance	216.68	136	0.027	0.998	43.3	30	0.05
	Eq. err variances	214.1	148	0.023	0.998	13.9	12	0.30
	Eq. latent variances	194.33	150	0.019	0.999	2.11	2	0.35
	Eq. correlations	190.4	154	0.017	0.999	4.42	4	0.35
	Diff. latent means	201.37	152	0.02	0.999	1.33	2	0.51
Wave 2	Baseline by mode design	185.92	90	0.035	0.997			
	Metric invariance	180.69	106	0.028	0.998	20.6	16	0.20
	Scalar invariance	219.44	136	0.026	0.998	49.1	30	0.02
	Free SF6a thresholds	210.93	133	0.026	0.998	40	27	0.05
	Eq. err variances	210.93	145	0.023	0.998	16	12	0.19
	Eq. latent variances	184.91	147	0.017	0.999	1.1	2	0.58
	Eq. correlations	184.25	151	0.016	0.999	5.69	4	0.22
	Diff. latent means	193.52	149	0.018	0.999	1.33	2	0.52
Wave 3	Baseline by mode design	211.97	90	0.049	0.998			
	Metric invariance	199.97	106	0.039	0.998	19.7	16	0.23
	Scalar invariance	230.23	136	0.035	0.998	45.7	30	0.03
	Free SF4b thresholds	223.48	133	0.034	0.998	38.6	27	0.07
	Eq. err variances	215.37	145	0.029	0.999	10.7	12	0.56
	Eq. latent variances	208.5	147	0.027	0.999	4.77	2	0.09
	Eq. correlations	194.98	151	0.023	0.999	3.08	4	0.54
	Diff. latent means	206.2	149	0.026	0.999	0.94	2	0.63
Wave 4	Baseline by mode design	210.04	90	0.05	0.996			
	Metric invariance	193.7	106	0.035	0.998	17	16	0.38
	Scalar invariance	205.37	136	0.031	0.998	32.3	30	0.35
	Eq. err variances	211.84	148	0.029	0.998	18	12	0.12
	Eq. latent variances	212.74	150	0.028	0.998	5.76	2	0.06
	Eq. correlations	211.41	154	0.027	0.998	7.79	4	0.10
	Diff. latent means	226.98	152	0.031	0.998	0.61	2	0.74

Gray background indicates decrease in the fit of the model.

Table 3.4: Mixed modes overestimate the threshold of SF6a compared to the single mode in wave two and underestimates the threshold of SF4b in wave three.

Wave	Threshold	Mixed mode	Single mode
Wave 2	SF6a$1	−1.718	−1.718
	SF6a$2	0.431	0.320
	SF6a$3	1.536	1.349
	SF6a$4	2.570	2.124
Wave 3	SF4b$1	−4.472	−4.472
	SF4b$2	−3.985	−3.254
	SF4b$3	−2.389	−2.231
	SF4b$4	−1.151	−1.122

and the observed scores (Millsap, 2012, Chapter 5). Thus, in the case SF6a in wave two we observed equality for the first threshold (indicated by "$1"), which is done in order to estimate the model (see Section 3.3), but for the rest we see that the mixed mode has larger values than the single mode. This indicates that even after controlling for true mental health, respondents in the mixed mode design tend to select more the first categories than those in the face to face single mode.

The differences found in the thresholds can be caused either by measurement, selection or an interaction of the two. Unfortunately they cannot be empirically disentangled using this research design. When considering measurement two main explanations appear: social desirability (Chen, 2008) and acquiescence. Due to the wording of the question, a higher score is equivalent to lower social desirability. As a result, if this is indeed the cause, then the mixed mode design, with the use of CATI, leads to less socially desirable answers. On the other hand, if acquiescence is the main cause, the systematic error is bigger in the mixed mode design. Alternatively, the difference may also mean that the CATI-CAPI sequential design tends to select more people who feel less often calm and peaceful (i.e., poorer mental health). Lastly, an interaction of the two explanations is also possible. For example, the mixed mode design may select fewer people who tend to respond in a socially desirable way.

In wave three, the thresholds of SF4b ('Did work less carefully') are significantly different between the two groups (Table 3.4). Once again, the respondents that took part in the mixed mode design in wave two tend to prefer the first answer categories

(worse health) compared to those in the single mode. Because the measurement was the same in this wave for both groups (i.e., CAPI), there are two possible explanations: attrition or panel conditioning. The latter is theoretically associated with increase reliability in time (e.g., Dillman, 2009), which would not explain differences in systematic error. As a result, the main theoretical explanation may be the different attrition patterns. This hypothesis is also supported by previous research (Lynn, 2013) which found different attrition patterns resulting from the mixed-mode design which disappears by wave 4.

Equivalence of latent growth models

Next, for each variable of the SF12, the LGM presented in Section 3.3 are tested using the $\Delta\chi^2$ method. For example, the Growth Model for SF6a (Table 3.5) has a good fit for the Baseline model, which does not impose any equality constraints between the two mode designs, RMSEA of 0.03 and CFI of 0.989. Imposing equal mean slope of change for the two groups does not significantly worsen the model ($\Delta\chi^2$ 1.04 with 1 df) while imposing equal variance of change, the equivalent of a random slope for time, leads to a significant $\Delta\chi^2$ (6.92 with 1 df). Lastly, imposing equal correlations between the intercept and the slopes does not reduce the fit significantly ($\Delta\chi^2$ 2.55 with 1 df).

The results indicate that four variables differ in their estimates of individual change (Table 3.5): SF6a ('Felt calm and peaceful'), SF6c ('Felt downhearted and depressed'), SF6b ('Lot of energy') and SF7 ('Social impact II') while the rest are the same (Table B.1). The first two are part of the same subdimension, mental health, while SF6b measures vitality and SF7 social functioning. All four are part of the mental dimension of the SF12 and differ in the same coefficient, the variance of the slope parameter (i.e., random effect for change in time).

A more detailed look indicates that the mixed mode design leads to the overestimation of individual change for all four variables: 0.116 versus 0.047 for SF6a, 0.078 versus 0.025 for SF6b, 0.108 versus 0.017 for SF6c and 0.134 versus 0.006 for

Table 3.5: For four out of the 12 items tested the mixed mode design has significantly different variance of the slope.

Variable	Model	χ^2	df	RMSEA	CFI	$\Delta\chi^2$	df	p
SF6a	Baseline by mode design	53.442	30	0.03	0.989			
	Equal mean of slope	51.64	31	0.027	0.991	1.04	1	0.31
	Equal variance of slope	58.717	32	0.031	0.988	6.92	1	0.01
	Equal correlation	58.343	33	0.029	0.988	2.55	1	0.11
SF6b	Baseline by mode design	94.013	30	0.049	0.985			
	Equal mean of slope	83.347	31	0.043	0.988	1.86	1	0.17
	Equal variance of slope	87.49	32	0.044	0.987	4.49	1	0.03
	Equal correlation	78.601	33	0.039	0.989	0.01	1	0.92
SF6c	Baseline by mode design	44.123	30	0.023	0.993			
	Equal mean of slope	42.992	31	0.021	0.994	0.69	1	0.41
	Equal variance of slope	51.625	32	0.026	0.991	8.98	1	0.00
	Equal correlation	48.285	33	0.023	0.993	1.43	1	0.23
SF7	Baseline by groups	51.677	30	0.028	0.99			
	Equal mean of slope	50.168	31	0.026	0.991	0.18	1	0.68
	Equal variance of slope	61.6	32	0.032	0.986	9.57	1	0.00
	Equal correlation	51.029	33	0.025	0.991	0	1	0.96

Gray background indicates decrease in the fit of the model.

SF7. A number of factors may explain the pattern. Firstly, the switch of mode may lead to changes that are not substantial (i.e., measurement noise) and, thus, biasing the estimates of change. Alternatively, the change of mode design can cause a decrease in panel conditioning, this, in turn, leading to a less stable change in time estimates. This seems less probable given Section 3.4 and previous research on this data. For example, Cernat (2015b) has shown that SF12, together with 20 other variables available in all the first four waves of the UKHLS-IP, have the same reliability in the face to face single mode design as in the mixed mode CATI-CAPI design. Lastly, non-response or attrition may cause a mode design effect that also impacts estimates of change. Previous research by Lynn (2013) has shown some effects of non-response in wave two on age, household type and car ownership, although these tend to disappear by wave four.

3.5 Conclusions and discussion

Overall the results show small differences between the two mode designs. When the modes are mixed (wave two of UKHLS-IP) significant differences are present only for one variable out of 12 (SF6a, 'Felt calm and peaceful'), with higher threshold for the mixed mode design. Two main explanations are put forward: measurement, through social desirability or acquiescence, and selection. Depending on the reference design, the systematic bias can be higher in either the mixed mode design (in case of acquiescence), or the single mode design (in case of social desirability). Alternatively, the mode design effect may be caused by non-response bias. The latter explanation is also partially supported by previous research (Lynn, 2013) and by the effect found in wave three.

Looking at the waves after the change to a mixed mode design was implemented shows, once again, either small or no differences. The only discrepancy appears in the threshold of a different variable, SF4b ('Did work less carefully'), in wave three. Here, because the same data collection procedure was used, two main explanation present themselves: attrition or panel conditioning. Theoretical and empirical results presented in the previous section support the former explanation.

The equivalence testing of the LGM shows that four of the SF12 variables have mode design effects in their estimates of individual change. For all four of them the same coefficient is biased in the same direction. It appears that for these items the mixed mode design overestimates variation of individual change. All four variables measure the same dimension, mental health, and use vague and subjective terms such as: calm, peaceful, a lot of energy or downhearted and depressed. One possible explanation can be that the mixed mode design adds extra noise that leads to overestimation of change in time. This may be especially the case for questions regarding subjective/attitudinal measures. Alternatively, the non-response bias observed in other studies may cause this pattern (Lynn, 2013).

The results of the study have a series of implications for surveys that plan to use mixed mode designs and for survey methodology more generally. On the one

hand, it appears that the mixed mode design (CATI-CAPI) has a small impact on the measurement (compared to CAPI). Nevertheless, when a mode design effect appears it may be persistent, although there is evidence that these tend to disappear after two waves (similar to the findings of Lynn, 2013).

Secondly, mixed mode designs can have an effect on estimates of individual change. While this effect was found in four out of the 12 variables analyzed, the differences can be up to six times larger in the mixed mode design. This change in mode design may lead to the overestimation of the variance of individual change in time (i.e., how different the change in time is between people). Attitudinal, subjective items may be especially prone to such effects.

Lastly, the paper has proposed two new ways of looking at mode design effects using equivalence testing in longitudinal data. Both of them can be used either with quasi-experimental designs or with other statistical methods that aim to separate selection and measurement. Equivalence testing with CFA has already proved useful in the mixed mode literature when applied to cross-sectional designs, such as those used by the European Social Survey mode experiments (Martin and Lynn, 2011b; Révilla, 2013).

As any study, the present one has a series of limitations. The first one refers to the design used by the UKHLS-IP. While it gives the opportunity to see the lasting effects of mixing modes, it is not a very common design. It is more likely that surveys will continue to use the mixed mode design after such a change takes place and not move back to a single mode design after one wave, as in the data used here. That being said there are examples of surveys that followed such a move. For example, the National Child Development Study will move back to a single mode after just one wave of using the mixed mode design.

Also, the paper does not aim to disentangle measurement and selection effects. While the use of randomization is used to associate the differences found to the mode design, other statistical models are needed to distinguish between measurement and selection into mode (e.g., Lugtig et al., 2011; Vannieuwenhuyze et al., 2012;

Schouten et al., 2013). Here only theoretical arguments and previous empirical work are explored as potential explanations. Additionally, the study analyses one type of scale (health related) with a particular type of mixed mode design (sequential) and a specific mix of modes (CATI and CAPI) in UK. As such, future research is needed to see if the findings are generalizable to other contexts.

Chapter 4

Estimation of Mode Effects in the Health and Retirement Study using Measurement Models

Abstract

Using multiple modes to collect data is becoming a standard practice in survey agencies. While this should save costs and decrease nonresponse error it may have detrimental effects on measurement quality. This can happen because different modes have distinct measurement biases which, when combined with selection effects, can increase the total survey error of a mixed-mode survey relative to a single mode approach. In this paper we use a quasi-experimental design from the Health and Retirement Study to compare the measurement quality of three scales between face-to-face, telephone and Web modes. Panel members were randomly assigned to receive a telephone survey or enhanced face-to-face survey in the 2010 core wave, while this was reversed in the 2012 core wave. In 2011, panelists with Internet access completed a Web survey containing selected questions from the core waves. We examine the responses from 3251 respondents who participated in all three waves, using latent models to identify measurement mode effects. Two of the scales, depression and physical activity, show systematic differences between interviewer administered modes (i.e., face-to-face and telephone) and the self-administered one (i.e., Web) while religiosity shows no differences of measurement between modes. Possible explanations are discussed.

4.1 Introduction

As surveys increasingly turn to mixed-mode designs, concerns about mode effects on measurement are being raised. And while mixed-mode strategies are often adopted for cost reasons, the trade-off in terms of measurement needs to be understood. This

is especially true of panel studies where a key focus is on measuring change over time and a necessary assumption is measurement invariance over waves of data collection (Cernat, 2015b,a). Much of the research on mode effects has involved cross-sectional designs, with subjects randomly assigned to one mode of data collection or another. This often makes it hard to disentangle selection effects (those who choose to respond in a particular mode) from measurement effects. Changing modes in a panel study may similarly confound true change with effects of mode (Cernat, 2015a). The optimal experimental design for disentangling selection and measurement effects while controlling for temporal change would involve randomly assigning subjects to different modes at different times (e.g., in a randomized cross-over design). Such designs (e.g., Gmel, 2000; Hays et al., 2009; Mavletova and Couper, 2013) are rare in large-scale panel studies because of their cost and effort to implement.

In this paper we exploit a design feature of the Health and Retirement Study (HRS) that was first introduced in the 2006 wave, in which a random half of the panel members are assigned to an enhanced face-to-face interview (which includes physical measurements and biomarker collection), while the rest are assigned to a telephone interview. In the next wave, these assignments are reversed so that each respondent gets the enhanced face-to-face interview every other wave (or every 4 years). In addition, those who have access to the Internet and are willing to do an online survey are invited to complete a Web survey in the "off-years" (i.e., the odd years between the even years of core data collection). While the content of these Internet surveys is typically focused on topics not asked on the core waves, or on experimental topics, in 2011 a set of questions was included in the Internet survey that is usually asked in the core, with the goal of exploring measurement effects of mode. We thus have a set of questions that are asked up to three times of the same respondents, once in a face-to-face interview, once by telephone (with the temporal order randomized) and once on the Internet (in between the other two waves). This design feature allows us to explore possible measurement differences across three modes for a select group of questions in the context of an ongoing representative

panel study.

In the sections that follow, we first review the literature on mode effects relevant to our study, then describe the modelling strategy we employ to isolate such mode effects. We then present the data and survey design in more detail, along with the specific hypotheses we test, before finally presenting the analyses and discussing the results.

4.2 Mode differences and previous research

Mode comparison studies - and hypotheses about causes for differences between modes - have a long history. Research on differences between face-to-face and telephone surveys date to the early introduction of the telephone mode (see Cannell et al., 1987; Groves, 1979; Herzog et al., 1983; Sykes and Collins, 1988; De Leeuw and van der Zouwen, 1988), but continues to receive attention (e.g., Béland and St-Pierre, 2008; Burton, 2012; Cernat, 2015b,a; Jäckle et al., 2006). Research comparing mode effects in Web surveys to interviewer-administered modes (telephone or face-to-face) is more recent (e.g., Chang and Krosnick, 2009; Dillman, 2005; Duffy et al., 2005; Fricker et al., 2005; Heerwegh, 2009). Given the many dimensions of mode (Couper, 2011), there are several mechanisms that could produce differences between modes in data collection. Our goal is not to attempt an exhaustive review of this literature, but to focus on two key aspects that are relevant for the items analysed here: interviewer administration versus self-administration and auditory versus visual presentation of survey questions.

One of the consistently found differences between interviewer-administered and self-administered surveys relates to social desirability bias, or the tendency to present oneself in a favourable light (see DeMaio, 1984). A number of studies have found higher reports of socially undesirable behaviors, attributes, or attitudes in self-administered surveys and lower reports of socially desirable ones (for reviews Groves et al., 2008; Tourangeau et al., 2000). These findings extend to Internet surveys (see, e.g., Heerwegh, 2009; Kreuter et al., 2008). While the differences between

face-to-face and telephone surveys are not as large, there is a general tendency for greater social desirability response bias on the telephone (see Holbrook et al., 2003).

Regarding the second feature of mode we explore, both face-to-face and telephone interviews involve interviewers, but may differ on the presentation of questions. Telephone is (by definition) aural, with the interviewer reading the question and response options to the respondent, who must keep this information in working memory while processing the question and formulating a response. Face-to-face surveys often involve the use of show cards, which display the response options to respondents, to minimize the cognitive burden of answering questions with several response options (see Lynn et al., 2012). HRS does not make use of show cards, so in this respect both the face-to-face survey and telephone survey can be viewed as primarily aural modes. In contrast, the Web is a primarily visual mode, with respondents reading survey questions on the Web page. This can lead to differential response order effects, with primacy effects (in which options presented first are selected more often) occurring in visual modes and recency effects (with later options selected more frequently) occurring in aural modes (see Krosnick and Alwin, 1987; Schwarz et al., 1992; Visser et al., 2000).

4.3 Measurement models and error

In order to evaluate data quality and relative bias we use the multiple items approach (Alwin, 2007). This implies the existence of a latent construct of interest, in our case continuous, that is measured with approximation by multiple observed variables. Models such as Confirmatory Factor Analysis or Item Response Theory use this approach, resulting in the following formulation:

$$y = \tau + \lambda\xi + \epsilon \tag{4.1}$$

where λ is the slope/loading or the strength of the relationship between the latent variable of interest, ξ, and the observed item, y. This can be considered an estimate

of reliability (Bollen, 1989), although it has a different meaning to that used in Classical Test Theory (Alwin, 2007; Lord and Novick, 1968). The random error, ϵ, is the complement of reliability and it can be easily calculated: $\epsilon = 1 - \lambda^2$. Lastly, τ represents the intercept, or the threshold when the observed variable is categorical, and can be interpreted as the conditional mean or probability of the observed items when the latent variable is 0. This is usually associated with systematic error (e.g., Chen, 2008).

This model has been further extended to a multi-group framework, enabling researchers to investigate relative bias between groups, such as sex, ethnicity or culture (Millsap, 2012) or, in our case, modes of data collection. This is not only an interesting methodological tool but it is also substantively important as differences in the measurement model across groups (called lack of equivalence or invariance) will bias comparisons of the latent variable.

The usual procedure in testing for equivalence of the measurement model across groups starts with the configural model (Meredith, 1993; Millsap, 2012; Steenkamp and Baumgartner, 1998). This implies that a model with the same structure is found in all the groups but no equality of coefficients is imposed. If this is found to have a good fit then the model is further restricted to assume equal loadings, λ, across groups. This is known as the metric equivalence (Steenkamp and Baumgartner, 1998). If this, in turn, fits the data, then a new model can be estimated which assumes that the loadings and the intercepts/thresholds, τ, are equal across groups. This model has been given different names by authors in this literature: scalar equivalence (Steenkamp and Baumgartner, 1998), strong factorial equivalence (Meredith, 1993) or first order equivalence (Millsap, 2012).

Using equivalence testing for estimating relative bias has become a standard procedure in cross-cultural research (e.g., Davidov et al., 2008; Van de Vijver, 2003) and it has also been implemented a number of times in the mixed-mode literature (e.g., Cernat, 2015a; Hox et al., 2015; Klausch et al., 2013). In this paper we combine the

use of this procedure with the quasi-experimental design of the data collection in order to estimate the effects of modes on measurement.

4.4 Research questions and theoretical expectations

The items chosen for inclusion in the 2011 Internet Survey were selected from among available core items (asked in 2010 and again in 2012) to test specific hypotheses related to mode effects. Here we concentrate on three scales that are measured by multiple items in all three waves: depression, physical activity and religiosity.

Generally the HRS does not contain very sensitive questions. Many of the questions that may be subject to social desirability effects are single-item (often yes/no) questions (e.g., alcohol use, seatbelt use, smoking status), that are not amenable to our analytic approach. But both the core and Internet surveys included the Center for Epidemiologic Studies Depression Scale (CES-D) measure of psychological distress, or symptoms of depression. This consists of a series of nine yes/no items, with three items reverse-scored, which will allow us to disentangle social desirability effects from response order effects. Depression measures have been found to be subject to mode-related social desirability effects (see, e.g., Moum, 1998), although Chan et al. (2004) suggest cognitive effects related to response order may be at work. Respondents who endorse four or more of the items are viewed as having depressive symptoms (Steffick, 2000). In addition, a three item physical activity index (frequency of mild, moderate, and vigorous exercise) was included in the Internet survey and core. Finally, we included a two item measure of religiosity (church attendance and importance of religion). As Presser and Stinson (1998) have documented, religious attendance is subject to social desirability bias associated with mode.

Based on the previous research, we expect more reports of depressive symptoms on the Web than in either interviewer-administered mode. Similarly, social desir-

ability biases should lead to lower reports of physical activity on the Web. However, this may be countered by response order effects (primacy on the Web), as the first option in each case indicates a higher level of activity (1 = more than once a week, 4 = hardly ever or never). Similarly, we would expect lower reports of religiosity on the Web, consistent with the social desirability hypothesis. But again, the first option for each of the two items is the high-frequency option (1 = more than once a week, 5 = not at all for religious service attendance; 1 = very important, 3 = not too important for importance of religion). In both cases, however, we expect the effect of social desirability to be stronger than that of primacy, so the overall net effect would be lower reports of physical activity and religiosity on the Web.

4.5 Data and design

Data for this study come from the Health and Retirement Study in the United States, a national panel study of men and women over the age of 50 that began in 1992. HRS conducts biennial interviews (in even-numbered years) with about 20,000 individuals. The sample is refreshed with a new cohort of individuals age 51-56 every six years (in 1998, 2004, 2010, etc.) to maintain representation of the population over age 50. Selected age-eligible respondents and their spouses of any age are interviewed. All baseline respondents (new cohorts interviewed for the first time) and persons 80 and older are assigned to a face-to-face interview, while the remainder are randomly assigned to either face-to-face (using computer assisted personal interviewing, or CAPI) or telephone (using computer assisted telephone interviewing, or CATI) mode. For panel (i.e., non-baseline) respondents under age 80 the mode assignment flips across waves (e.g., from telephone in 2010 to face-to-face in 2012 or vice versa). Response rates for the core interview have ranged from 52 to 81% at baseline and from 87 to 89% at each follow-up wave.

In addition to the biennial Core interview, HRS also conducts a number of supplemental studies, mainly in the form of mail and Internet surveys that are conducted in the off-year between interview waves. The Internet survey has been ongoing since

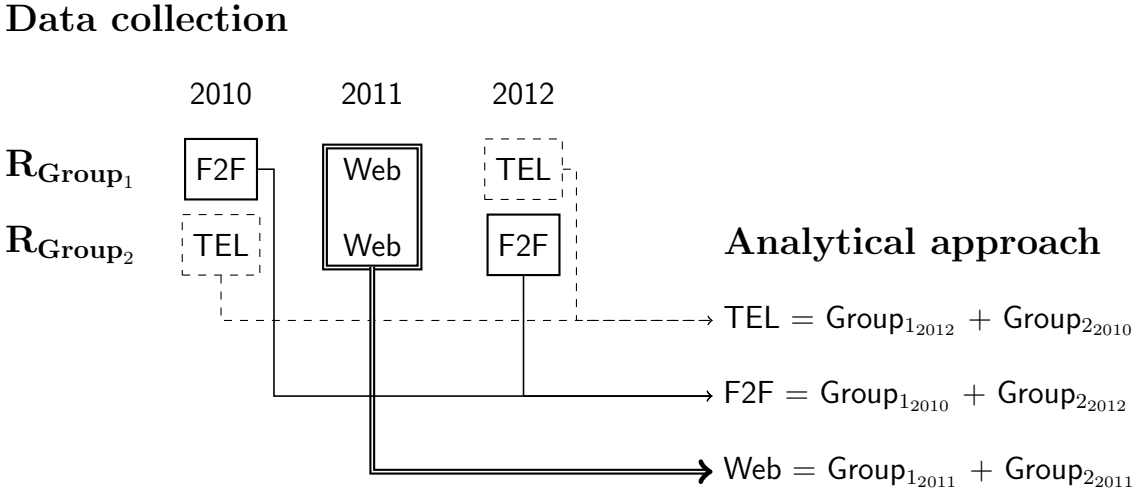
2001 and is administered to respondents who report in their core interview that they have Internet access. The 2011 HRS Internet survey included a number of items to explore possible mode effects, repeating measures that were asked in the 2010 and 2012 core interviews. The response rate for the 2011 Internet survey was 81%. A total of 3251 respondents who were subject to the random mode rotation completed all three surveys and comprise our analysis sample. Of these, 1583 were assigned to a telephone interview and 1668 to face-to-face in 2010. This sub-group of respondents represents 70.8% of participants in the 2011 Web survey and 14.8%/15.8% of the 2010/2012 HRS respondents.

The link between data collection and our analytical approach is shown in Figure 4.1. It can be seen that in 2010 two groups were randomly allocated to either face-to-face (Group 1) or telephone (Group 2). The order was reversed in 2012. In the year between these two waves all selected respondents answered a Web survey. On the right side of the Figure we can see how this translates into our analytical groups. Thus, each individual answers in all three waves. We also observe how this design partially avoids confounding time with mode. This is only partial as all Web responses come from the 2011 wave. If there are time specific or non-linear learning effects then these may bias interviewer vs. Web comparisons. This potential confounding is partially solved by the statistical approach used here which lets the latent, or "true", variables of interest be different across modes. Additionally, the analysis was rerun using the mode of interview in wave 2010 as a control variable. This will be a sensitivity check for the impact of the order in which the modes of interview were received.

Data management

The analysis uses a balanced panel of the respondents that took part in the 2010, 2011, and 2012 waves of the HRS. The mode variable used reflects the mode in which the interview was assigned. As noted previously, mode for the core interview was randomly assigned for panel respondents under age 80, with roughly half being

Figure 4.1: The link between the quasi-experimental data collection design and analysis strategy



assigned to telephone and half to face-to-face. Although interviewers make every attempt to complete the interview in the assigned mode, in some circumstances respondents are allowed to switch modes. Only a small proportion of respondents in our sample did not complete their interview in the assigned mode (3.1% in 2010 and 4.8% in 2012). The most common switch was from face-to-face to telephone, though some respondents also switched from telephone to face-to-face. Additionally, there are respondents that answered using the same mode in both 2010 and 2012: 155 (4.8%) answered by telephone in both 2010 and 2012 waves while 92 (2.8%) answered by face-to-face in both waves. As a sensitivity analysis all the models were rerun on the more restricted sample that includes only people that actually switched modes between 2010 and 2012. Missing data was low for the items we examine, the highest being 1.3% for the "Had a lot of energy" item (details can be found in the Annex). The analysis uses Full Information Maximum Likelihood (FIML) to deal with missing data and assumes missingness at random (MAR) given the measurement model (Enders, 2010).

Analytical approach

Using the data and the statistical method presented above we test a series of nested models to identify different types of measurement mode effects. The sequence will

distinguish between random error (evaluated based on the loadings with metric equivalence) and systematic error (evaluated based on thresholds with scalar equivalence) and between modes: telephone (TEL) versus face-to-face (FTF) and interviewer versus self-administered (FTF and TEL vs. Web). From these theoretical comparisons stem the five (cumulative) models tested:

- **Configural** (structure is the same in all modes, no equality constraints);
- **Interviewer metric equivalence**: the same loadings in FTF and TEL;
- **Full metric equivalence**: FTF, TEL and Web have the same loadings;
- **Interviewer scalar equivalence**: the same thresholds in FTF and TEL;
- **Full scalar equivalence**: the same thresholds in FTF, TEL and Web.

This sequence of models reflects our theoretical hypotheses regarding the mode impact on measurement. We expect FTF and TEL to be more similar as both of them are mainly aural and involve communication with an interviewer. Nevertheless some differences are expected due to higher social desirability and faster pace in TEL (Holbrook et al., 2003). On the other hand, we expect the Web to show the biggest differences in relative systematic bias. Firstly, it is self-administered, as such we expect smaller social desirability effects. Secondly, it is mainly visual, which might lead to primacy effects.

It should be noted that in all these models no assumption is made about the equality of the latent variables (either mean or variance) across modes. Thus, any learning or maturation which might appear and is not controlled for by our quasi-experimental design are expected to appear as differences in the latent variable.

To estimate the models we use Maximum Likelihood Robust estimation as implemented in Mplus 7.2. All the observed variables are considered categorical while the latent variable is modelled as continuous. As such, thresholds are calculated (number of thresholds is one less than the number of categories) and compared across modes in

order to estimate systematic error. This can be viewed either as a categorical Multi-Group Confirmatory Factor Analysis or as an IRT model (Kankaraš and Moors, 2010; Millsap, 2012). Models are compared by using a corrected score of the $\Delta\chi^2$. This is calculated by the difference in χ^2 of two nested models. The degree of freedom of the test is the difference in degrees of freedoms between the models compared. A correction is applied to the score in order to take into account the Maximum Likelihood Robust estimation (Satorra and Bentler, 2001)[1]. The Akaike Information Criteria (AICs) are also reported. This is an indicator of relative fit based on the log-likelihood of a model that 'penalizes' for lack of parsimony. A smaller AIC implies a better fitting model.

4.6 Results

Depression scale

The first scale analysed using the procedure presented above is the CES-D, which estimates depressive symptoms. An underling continuous latent variable was modelled with 9 dichotomous observed items (frequencies can be found in the Annex). The first model, Configural, assumes that the structure of the measurement model is the same across modes (e.g., no correlated errors in one of the modes) but does not impose equality constraints on the coefficients across modes. The second model, Interviewer metric equivalence, assumes equal loadings, or reliability, across TEL and FTF. Table 4.1 shows that the Interviewer metric equivalence model should be selected as it does not fit significantly worse than the Configural model even if it more restrictive (p-value of 0.85 and AIC is smaller). Similarly, the third model, which assumes equal loadings across all three modes, fits the data well, indicating that Web does not differ in reliability compared with TEL and FTF (p-value of 0.83 and AIC is smaller). Looking at the mode effects on systematic measurement we find no differences between TEL and FTF (p-value of 0.72 and AIC smaller);

[1]See http://www.statmodel.com/chidiff.shtml for explanation and an example.

Table 4.1: Equivalence testing of the CES-D and thresholds for interviewer and Web.

Model	χ^2	df	$\Delta\chi^2$	p-value	AIC
Configural	3308.315	1446			77554
Interviewer metric equivalence	3315.851	1454	4.04	0.85	77542
Full metric equivalence	3357.293	1463	5.03	0.83	77531
Interviewer scalar equivalence	3355.668	1472	6.18	0.72	77519
Full scalar equivalence	3379.57	1478	99.39	0.00	77602

Threshold	Interviewer	Web
Depressed	6.58	5.62
Effort	3.50	3.25
Sleep	1.39	1.30
Happy	-4.54	-3.93
Lonely	3.39	3.13
Life	-6.54	-5.07
Sad	3.73	3.64
Not get going	3.11	2.76
Energy	-0.59	-0.39

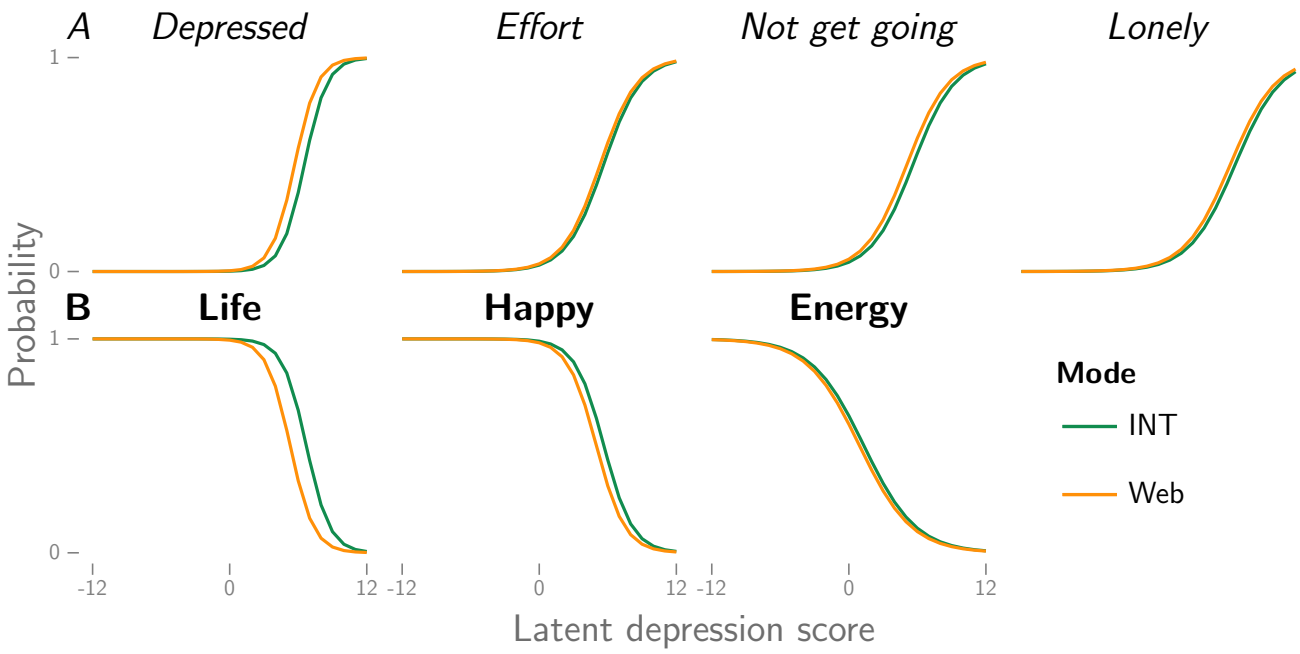
however these two modes are systematically different from Web (p-value of 0.00 and AIC is larger). This indicates that the relative measurement quality is the same across modes with the exception of systematic errors between interviewer modes and Web. These results are consistent with the sensitivity analysis done using only the respondents who changed the modes in 2010-2012 and/or controlling for the mode order (not shown).

We are able to further investigate the differences indicated by these analyses. The lower part of Table 4.1 shows the thresholds for the two interviewer modes and those from the Web responses (from the Interviewer scalar model). Further testing has shown that all the differences in thresholds are reliable with the exception of the 'Sleep' and 'Sad' variables. When we free (i.e., allow to be different) all the thresholds with the exception of these two the model is not significantly worse than Interviewer scalar equivalence ($\Delta\chi^2_2 = 1.81$, p-value of 0.40).

Because the observed variables are dichotomies (no/yes) the model estimates one threshold for each item. A large number on the threshold means that there are more people answering the first category (in this case 0 = no) after controlling for their true depression score. Differences across groups in thresholds imply relative

systematic measurement differences. The results show that for all the negatively worded items that are significantly different ('Depressed', 'Effort', 'Lonely', 'Not get going'; 1 = yes = more depression) the thresholds are lower for the Web while for all positively worded items ('Happy', 'Life' and 'Energy') the thresholds are higher (more no's). This means that even after controlling for their latent score, responses in the Web mode indicated higher depression levels than those from TEL and FTF. The most plausible explanation for this pattern is higher social desirability bias in the interviewer modes. Because the scale includes both positively and negatively worded items, response order effects (primacy/recency) can be ruled out.

Figure 4.2: Item characteristic curves for "Yes" in the significantly non-equivalent CES-D items, interviewer vs. Web.



To make this pattern clearer we have calculated and plotted the Item Characteristic Curve (ICC) for all significant differences (Figure 4.2). This plots the probability of selecting a certain category (y axis), in this case saying "Yes", based on the latent score of interest (x axis), depression. The verticality of the line is influenced by the discrimination or loading of the item. The flatter it is the less information it gives. The horizontal position indicates difficulty or the threshold and tells us at what levels of the latent variable does the item give information. In

Figure 4.2, for example, saying "Yes" to the 'Depressed' item has a high level of discrimination, quite vertical, and is also is an indicator of a relatively high level of latent depression. What is interesting for us is how this curve is different between interviewer and Web modes. We can see that the angle of the curve is the same, due to the equal loadings, but the horizontal position is different. So, for the same level of latent depression respondents are more likely to say "Yes" to the 'Depressed' item on the Web than in a interviewer administered survey (Figure 4.2A). The opposite is true for positively worded items such as 'Happy' (Figure 4.2B). In this case one is more likely to answer "Yes" in interviewer modes given the same level of latent depression. This pattern is consistent with social desirability.

In order to provide a sense of the differences between the two types of modes we can look at the variables that have the biggest and those that have the smallest significant differences (as seen in Figure 4.2). Because the predicted probabilities depend on the score of the latent variable we are going to choose values on this scale that highlights the biggest mode difference for each variable/category. For example, in the case of the 'Life' variable for a score of 6 (range -12 to 12) on the latent depression scale respondents in the interviewer-administered modes have a predicted probability of 67% to say 'Yes' compared to 34% for Web responses. We believe that this would be an substantially important difference in most applied research. At the other extreme this differences is approximately 5% for the 'Effort' item (56% for interviewer surveys compared 60% for Web).

Activity scale

The second scale we analyse measures physical activity. This is based on three observed variables that ask about the frequency of different types of activities: mild, moderate and vigorous. Table 4.2 shows that the loadings, or reliabilities, are equal across all three modes, indicated by the fact that the second and third models are not significantly worse than the previous ones (p-values of 0.37 and 0.12, both AICs are smaller). On the other hand, the thresholds, or relative systematic error,

are the same between face-to-face and telephone (p-value of 0.98 and AIC smaller) but these two are systematically different from Web (p-value of 0.00 and AIC is larger). This implies that the level of physical activity appears to be measured systematically differently in face-to-face and telephone, on one hand, and Web, on the other. Further testing showed that only part of these thresholds is significantly different. Thus, when comparing interviewer modes with Web the third threshold for all the variables and the first threshold of "Mild activity" are significant different ($\Delta\chi^2_5 = 4.13$, p-value of 0.53 when these are freed). These findings were replicated in our sensitivity analyses when we control for mode order effects and/or restricted the sample only to people that changed mode of interview.

The different levels of the thresholds can be seen in the lower part of Table 4.2 and their effects on the ICC's are apparent in Figure 4.3. We see that for all three variables Web respondents are less likely to choose the last category, 'Hardly ever or never', and are more likely to choose 'One to three times a month' for "Mild" at the same levels of latent physical activity. These differences are moderate to large as can be seen when we analyse the predicted probabilities of selecting a category for different scores on the latent physical activity scale (range from -3.5 to 3.5). For example, looking at the predicted probability of selecting the 'Hardly ever or never' category we find a difference of approximately 9 percentage points for the "Mild" and "Vigorous" items (47% versus 38% and 57% versus 46% at a score of 3.5 and of 0.5 on the latent physical activity scale for interviewer versus Web responses). The biggest difference can be found on the probability of answering the same category for the "Moderate" item at a level of 1.5 on the latent physical activity variable: 81% for interviewer answers versus 31% in Web interviews.

Such a pattern can be explained both by primacy/recency effects, Web respondents being more likely to choose the first categories while in the auditory modes the last ones, and higher social desirability bias when answering using Web. While our initial expectation was that social desirability would be stronger in the interviewer modes this does not appear to be the case. The opposite can be observed in our

Table 4.2: Equivalence testing of the activity scale and thresholds for TEL, FTF and Web.

Model	χ^2	df	$\Delta\chi^2$	p-value	AIC
Configural	1251.302	153			80920
Interviewer metric equivalence	1241.365	155	1.97	0.37	80917
Full metric equivalence	1224.226	157	4.17	0.12	80916
Interviewer scalar equivalence	1226.26	166	2.45	0.98	80900
Full scalar equivalence	1330.319	175	205.36	0.00	91088

Threshold	Interviewer	Web
Mild1	0.857	1.027
Mild2	2.575	2.498
Mild3	3.636	3.949
Moderate1	1.809	1.853
Moderate2	5.856	5.972
Moderate3	9.445	11.787
Vigorous1	-1.014	-0.962
Vigorous2	-0.335	-0.252
Vigorous3	0.282	0.774

data as interviewer modes systematically under-report physical activity compared with Web answers. Although we cannot disentangle primacy/recency from social desirability for this scale, higher recency levels in interviewer modes seems the most plausible theoretical explanation for the observed pattern. The absence of social desirability effects could be explained by the fact that the fitness of respondents is an observable attribute which may lower social desirability bias in interviewer modes (see Tourangeau et al., 2000, for overview).

Religiosity

The third scale tested measures religiosity using two indicators: importance of religion (three answer categories) and religious service attendance (five answer categories). Here we expect differences both between telephone and face-to-face (Holbrook et al., 2003; Presser and Stinson, 1998) and between these two and the Web answers. The main potential cause for such differences would be social desirability.

Both the $\Delta\chi^2$ and the AIC indicate that random and systematic errors are

Figure 4.3: Item characteristic curves for significantly non-equivalent activity variables/categories, interviewer vs. Web.

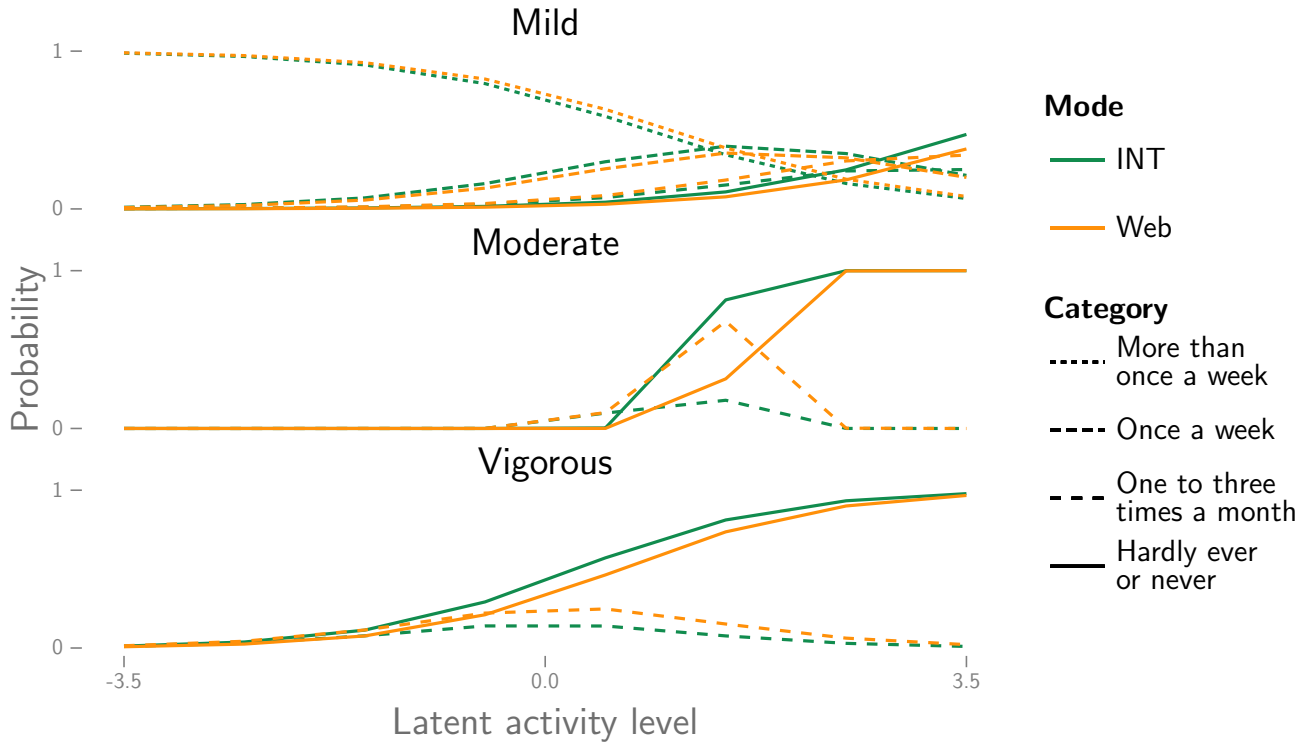


Table 4.3: Equivalence testing of the religiosity scale and thresholds for TEL and FTF.

Model	χ^2	df	$\Delta\chi^2$	p-value	AIC
Configural	287.947	18			66085
Interviewer metric equivalence	290.726	19	0.07	0.80	66083
Full metric equivalence	289.92	20	1.96	0.16	66081
Interviewer scalar equivalence	293.692	26	4.58	0.60	66074
Full scalar equivalence	302.044	32	6.11	0.41	66068

the same across the three modes (none of the models are significantly different in Table 4.3). This indicates that, unlike our theoretical expectation and the two previous scales, the measurement quality of this scale is the same across modes. The sensitivity analysis, controlling for mode order and/or analysing only people who changed modes, support these conclusions as no significant difference between modes was found.

4.7 Conclusions

In this paper we used a quasi-experimental design implemented in the 2010-2012 waves of the Health and Retirement Study to estimate mode effects on measure-

ment. Using latent measurement models we compared random and systematic error on three scales: depression, physical activity and religiosity. The results partially support our hypotheses regarding mode effects.

Previous literature regarding mode effects on measurement has consistently found social desirability bias as an important source of differences. This was partially replicated in our analyses. The CES-D depression scale enabled us to separate social desirability from primacy/recency effects. We show that responses collected in interviewer modes are consistently influenced by social desirability compared to Web, this resulting in higher observed levels of depression even after controlling for the latent level of depression. On the other hand, the religiosity scale did not present any mode differences due to social desirability. Another possible cause for mode effects put forward was primacy/recency effects. This was partially supported by our results as the Web respondents report higher levels of physical activity, consistent with higher recency effects in aural modes (i.e., telephone and face-to-face without showcards).

As in all research our study has several limitations. Firstly, the respondents included in the analyses is a sub-group of a representative sample of the population over 50 that have access to the Internet and who participated in three waves of a longitudinal study. Secondly, our study looks only at three scales. Different patterns may be expected for other topics and other types of response scales.

Nonetheless, these findings have important implications for survey methodology, although they are mostly in tune with a growing body of literature on the topic. First of all, the biggest differences we found were between interviewer and self-administered modes. Our hypothesised reasons, social desirability and recency/primacy, finds some support in our analyses. Secondly, we saw that two out of the three scales lack equivalence in the systematic part of the measurement model across interviewer/Web modes. This implies that using a mixed-mode design may lead to lower levels of equivalence which, when combined with selection effects, could bias substantive results. Thus, a combination of improvements in

design that would minimise mode measurement effects, and statistical approaches to correct for these, such as the use of instrumental variables or of the front-door approach (Vannieuwenhuyze et al., 2014; Cernat, 2015c), are advised. The front-door approach has been recently proposed as an alternative that aims to control for causes of mode measurement effects in order to estimate selection into modes. The type of analyses carried out in this paper would be especially useful when using such a statistical approach. Finally, in tune with other research on the topic, we caution against mixing interviewer and self-administered modes, when possible, and encourage study designs that allow for the evaluation of mode effects across a range of topics and indicators.

Chapter 5

The role of email contact in determining response rates and mode of participation in a mixed mode design

Abstract

This paper is concerned with the extent to which the propensity to participate in a web - face-to-face sequential mixed-mode survey is influenced by the ability to contact sample members by email in addition to mail. In panel surveys, researchers have the opportunity at each wave to ask for an email address, but there is little evidence regarding the value of doing so. This makes it difficult to decide what efforts should be made to collect such information and how to subsequently use it efficiently. Using evidence from a randomised experiment within a large national survey, we find that using a respondent-supplied email address to send additional survey invites and reminders does not affect survey response rate compared to using mailed invites and reminders alone, but is associated with an increased proportion of responses by web rather than face-to-face and, hence, lower survey costs. We find no evidence that these results depend on time in sample or time since the email was provided.

5.1 Background

In longitudinal surveys, researchers can ask sample members to provide their email address in order to contact them at subsequent waves. A similar opportunity may also arise in some types of one-time web surveys such as visitor surveys, where

visitors may be handed a card or letter asking them to go online and complete a survey. They could at the same time be asked to supply an email address. However, asking sample members to provide an email address is not cost-free. The request may be seen as intrusive and the information as sensitive and private. This could impact negatively on the propensity to participate in the survey (though Bandilla et al. (2014), found no effect of asking for an email address on participation in a follow-up survey). Furthermore, resources are required to capture, clean and manage the collected email addresses. Researchers should therefore be reassured of the value of asking for an email address before doing so. The focus of this article is sequential mixed mode surveys in which the first phase (mode) is web and the second is face-to-face interviewing and in which the first communication with sample members is by mail and includes an invitation to participate online. This is a common type of design (Lynn, 2013; Millar and Dillman, 2011) and is therefore a context of interest to many researchers. Moreover, we are concerned with the use of email to make additional contacts (invitation or reminders) during the first (web) phase of field work, not to substitute mail contacts.

There are two potential advantages of being able to contact sample members by email. First, it could increase the overall propensity for survey participation. Second, it might reduce data collection costs if it results in a higher proportion of response by web mode rather than an interviewer-administered mode. The mechanisms that could bring about each of these two effects are discussed in the next sections. Aside from response propensity and cost, speed of response can also be an advantage of email contact (Mehta and Sivadas, 1995; Schaefer and Dillman, 1998), but this consideration only applies to single-mode web surveys in which all sample members can be contacted by email. In mixed mode surveys, the completion of field work generally must await the slowest mode, so speed of response is not considered further in this article.

To the knowledge of the authors no study has investigated the effect of additional email contact on response propensity in either a mixed-mode or longitudinal context.

As a result, this article addresses an important methodological question that has yet to be tackled. Furthermore, the use of a nationally-representative sample and a quasi-experimental design provide a strong basis for inference and a context from which a degree of generalisability can be assumed.

Email contact and response propensity

There are at least two mechanisms through which additional email communications could increase response propensity. First, emails could reduce the risk of failing to make contact with the sample member. Non-contact is one of the major components of survey nonresponse (Groves and Couper, 1998) and the probability of it occurring depends on the number, nature, and timing of contact attempts (Lynn, 2008). Email communications are very different in nature to mail communications in a number of ways that are relevant to contact propensity. They tend to arrive in a personal inbox, checked only by the intended recipient, whereas mail is delivered to a letterbox that may be shared by other residents of the address. Consequently, mail can be removed by another person before the intended recipient sees it, whereas email generally cannot. Also, most people have opportunities to check their email inbox several times a day and can do so from multiple locations, whereas checking a mail box requires physical presence and may not be something that is done often. For these reasons, additional email communications should tend to increase contact propensity and result in the communication being seen by some sample members who would not have seen it had it only been sent by regular mail.

The second mechanism by which additional email communications could increase response propensity is through reduction of the burden of participation. Respondent burden (Bradburn et al., 1978; Sharp and Frankel, 1983) is conceptualised as encompassing the time it takes to perform survey tasks and the associated disruption to the respondent's other activities (as well as other features such as cognitive effort and embarrassment). Greater burden can reduce the probability that a sample member will initiate, or continue with, survey tasks. Sending a survey invitation by

email enables the recipient to participate by simply clicking on a link while already online, whereas if the invitation is received by regular mail the recipient must retain the letter until it is convenient to go online and must then type in a URL and enter a passcode. The latter clearly takes more time and requires more effort; the increased burden could reduce participation propensity (Millar and Dillman, 2011). Millar and Dillman (2011) found that adding two email contacts in a single-mode cross-sectional web survey which otherwise involved three mail contacts significantly increased response rate, though their study was of undergraduate students, all of whom had email addresses and were assumed to be web users.

Several other studies have examined aspects of the use of email contacts in the single-mode web context, but none of these studies assessed the effect of email contacts additional to mail contacts. The effect of substituting email contacts for mail contacts was tested by Porter and Whitcomb (2007), Millar and Dillman (2011) and Kaplowitz et al. (2012), all of whom found no effect. Kaplowitz et al. (2004) compared different combinations of email and mail contacts, but all treatments included an email contact. Bandilla et al. (2012) found higher response rates with mail rather than email invitations (in the absence of a mailed prenotification letter). Bosnjak et al. (2008) found higher response rates with email invitations rather than SMS invitations. A meta-analysis carried out by Manfreda et al. (2008) found that web surveys achieved a higher response rate when the invitation was delivered by email rather than mail, but they too did not assess the marginal effect of email contacts additional to mail contacts. Muñoz-Leiva et al. (2009) found that additional email reminders could increase response rates when previous contacts had also been by email, but did not compare treatments that involved mail contacts. Bosnjak et al. (2008) compared mode of prenotification, but not of invitation or reminders.

Email contact and mode of response

In a sequential mixed mode design where sample members are first invited to complete the survey online and subsequently approached for interview only if the online survey has not been completed, additional email contacts could increase the propensity to complete the survey online, even if overall participation propensity (as discussed in the previous section) is not affected. In other words, conditional on participation, respondents may be more likely to participate in web mode rather than using an interviewer mode. This is a desirable outcome for the researcher as data collection costs are reduced. The mechanisms through which this shift in the distribution of mode of participation could occur are essentially the same ones outlined above: the email invitation may increase the probability of the sample member being aware of the invitation (contact) or may make online participation easier (burden). Whether the outcome of these mechanisms is to increase the overall participation propensity or to increase the proportion of responses that are made online will depend on the extent to which sample members who only participate online as a result of the email communications are people who otherwise would not have participated at all (overall participation propensity) or are people who otherwise would have participated by interviewer-administered mode in the second phase of the field work (proportion of responses made online).

Moderating factors

Any effects of email communications, via the mechanisms of increased contact propensity or reduced burden, may be moderated by other factors. Three types of factors can be identified: socio-demographic characteristics, reactions to the request to provide an email address, and survey characteristics. A wide range of socio-demographic characteristics of survey sample members have been found to moderate the effectiveness of survey design features intended to increase participation propensity. In the longitudinal survey context, reviews of such effects can be found in Watson and Wooden (2009) and Uhrig (2008). Non-response theory does

not posit that these characteristics have a direct causal effect, but rather that they act as markers for variations in at-home patterns, time availability, psychological dispositions, and relevant attitudes (Groves and Couper, 1998; Groves et al., 2000). Knowledge of the moderating effects of socio-demographic characteristics can be useful to researchers implementing longitudinal surveys as design features can then be targeted at subgroups for whom they are expected to be effective (Lynn, 2014b).

There are two aspects of the sample member's reaction to requests for email addresses that can be of operational interest to longitudinal survey researchers. The first is how recently an email address was provided. Any moderating effect of this on the effect of email communications would have implications for how frequently researchers should ask sample members to provide an (updated) email address and/or for which sample members should be sent email communications (only those who provided/confirmed an email address relatively recently, or also those who provided an email address longer ago). The second aspect of interest is the reaction of other household members to the request to provide an email address. (This applies only to surveys that collect data from multiple members of a household.) Other household members may influence a sample member's survey participation decision and this may be particularly likely in the case of spouses and partners, whose relationship will tend to be the closest. For example, a person who has not themselves provided an email address and therefore cannot be sent a survey invitation by email may nevertheless be influenced by the inclusion of email invitations in the survey design if their partner receives one and tells them about it.

Another survey characteristic pertinent to decisions about design features for longitudinal surveys is time spent in the sample (Kalton and Citro, 1995). Some features may be more effective for recently-joined sample members (or in the early waves of a survey, in the case of a fixed sample), while others may work better amongst long-term sample members (or in the later waves of a survey). For example, the need to explain the content and relevance of the survey may be greater amongst recently-joined sample members (Lynn, 2014a), while certain tracking procedures

may be more effective amongst long-term sample members (Couper and Ofstedal, 2009).

5.2 Research questions

This article is concerned with the context of mixed-mode longitudinal surveys in which the first mode in a sequential design is web. Our interest is in the effectiveness of requesting email addresses at each wave and subsequently using the collected addresses to provide an additional channel of communication for survey invitation letters and reminders. Effectiveness is defined by two outcomes: participation propensity at the wave involving email communication, and the proportion of responses that are obtained online rather than through interviewer administration. Additionally, there is an interest in identifying moderators of any effects that are found.

Our research questions are therefore:

a) Does the use of email for additional contact attempts affect the overall propensity of sample members to participate in the survey?

b) Are the effects on participating in the mixed mode design influenced by the partner's provision of email or by when the email was provided? Moreover, are any of these effects moderated by characteristics of sample members or by time in sample?

c) Does the use of email affect the conditional propensity of sample members to participate in web mode rather than interviewer-administered mode?

d) Are the effects on participating by web influenced by the partner's provision of email, by when the email was provided and are any effects moderated by characteristics of sample members or by time in sample?

5.3 Study design

We use data from wave 5 of the Innovation Panel component of Understanding Society, the UK Household Longitudinal Study (UKHLS-IP). The UKHLS-IP (Uhrig, 2011) is designed specifically for methodological development and testing, primarily to inform the design of the main UKHLS, which is the UK's largest social science research resource investment (Buck and McFall, 2012; Hobcraft and Sacker, 2012). It is based on a stratified, clustered, probability sample of residential addresses in Great Britain (Lynn, 2009). All current residents at sample addresses in April to June 2008, when interviewers carried out wave 1 of the survey, were designated sample members and were followed up for subsequent waves at approximately one-year intervals. A refreshment sample, selected through the same design, was added at wave 4. At each wave, data are collected from all adult members of the household, even though not all such people are themselves sample members[1]. At each wave, respondents are asked to provide a range of contact information, including email addresses. Waves 1, 3 and 4 involved single-mode Computer Assisted Personal Interview (CAPI) data collection, while wave 2 had an experimental Computer Assisted Telephone Interview - CAPI mixed mode design (Lynn, 2013).

Field work for wave 5 took place in May to July 2012. A random two-thirds of sample households were allocated to a web-CAPI sequential mixed-mode design, while the other one-third was administered single-mode CAPI. This randomised allocation to mode treatment is what allows us to identify the effect of email communications on participation rates. How we do this is discussed in the next section. In the mixed-mode treatment, each sample member aged 16 or over was sent a letter with an unconditional incentive, inviting them to take part by web. The incentive took the form of a voucher for £5, £10, £20 or £30, each sample member having

[1]This study is concerned with response by adults (persons aged 16 or over) to the individual interview, which averages around 35 minutes. The UKHLS-IP also involves a self-completion questionnaire for children aged 10 to 15 and a household enumeration and questionnaire, which averages around 12 minutes and is completed by one adult in each household. We do not consider here response to either of those instruments.

Table 5.1: Survey contact sequence for each sample group

Treatment	Email address	Day 1: Mail invite	Day 2: Email invite	Day 5: Email reminder	Day 8: Email reminder	Day 14: Mail reminder	Day 15-35: CAPI fieldwork	N
Single-mode CAPI	Yes or No						✓	857
Mixed mode Web-CAPI	Yes	✓	✓	✓	✓	✓	✓	889
	No	✓				✓	✓	776

been randomly allocated to one of four experimental groups [2]. The letter included the URL and a unique user ID, to be entered on the welcome screen. A version of the letter was additionally sent by email to all sample members for whom an email address was known (just over half of the sample). For people who had indicated at previous waves that they do not use the internet regularly for personal use (20% of respondents), the letter informed them that they would have an opportunity to do the survey with an interviewer. Up to two email reminders were sent at three-day intervals. Sample members who had not completed the web interview after two weeks were sent a mail reminder and interviewers then started visiting to attempt CAPI interviews. The interviewer visits began in the same week that the reminder letter would have been received in order to constrain the overall field work period. The web survey remained open throughout the fieldwork period.

In the single-mode CAPI treatment, each sample member was sent a letter with an unconditional incentive, explaining that an interviewer would soon visit their address. The proportion of incentives of each value was identical to that for the mixed-mode group and the design and content of the letter was identical aside from the paragraph that mentioned an interviewer visit instead of inviting online participation. Copies of the invitation and reminder letters for both treatment groups are included in the annex. The contact sequence for each sample group is summarised in Table 5.1.

The present study is based on sample members issued to the field for wave 5

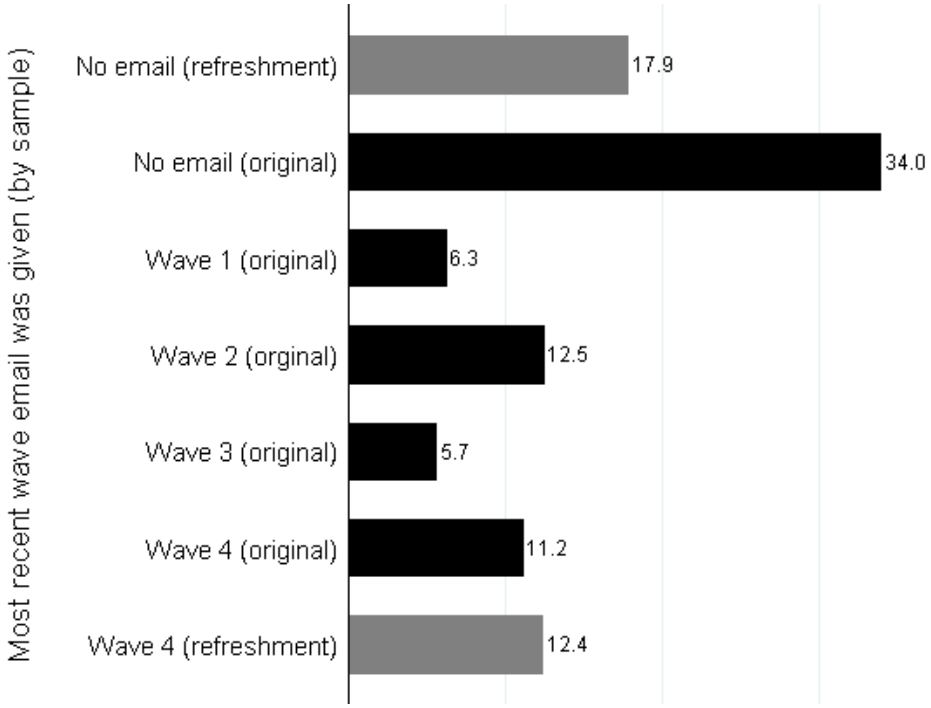
[2]The four incentive treatment groups are orthogonal to the two mode treatment groups, so there is no confounding of these two effects.

(n=2,522). The outcomes of the wave 5 fieldwork are our dependent variables of interest. As outlined above, the sample issued at wave 5 consisted of two components: the original sample, participating for the fifth time, and the refreshment sample, participating for the second time. Estimated response rate to the wave 1 enumeration was 60.9% (AAPOR RR1). Of all persons aged 16 or over enumerated at wave 1 and not known to have become ineligible prior to wave 5, 66.0% (1,819 persons) were issued to the field for wave 5, the rest having been lost due to a failure to trace following a move, persistent non-contact, or refusal. Estimated response rate to the wave 4 enumeration of the refreshment sample was 61.4% (AAPOR RR1), all of whom were issued at wave 5 (887 persons aged 16 or over). The present study is therefore based on around 40.2% of original sample members and 61.4% of refreshment sample members. This corresponds to 45.7% of all sample members.

5.4 Data and methods

Our dependent variables are indicators of whether the sample member completed the individual interview at wave 5 and, if so (for the mixed-mode sample), whether they completed it in web mode or by CAPI. Our key independent variables are dichotomous, taking the value 1 if a characteristic or design feature applies and 0 otherwise. **Mode treatment** indicates whether the sample member was allocated to the mixed-mode treatment rather than the single-mode CAPI treatment; **Time in sample** indicates membership of the original sample rather than the wave 4 refreshment sample. **Email** indicates whether an email address was supplied by the sample member prior to wave 5. Note that this is independent of **Mode treatment**: the request to provide an email address was made of all sample members at waves 1 to 4 without knowledge of the mode treatment to which they would be assigned at wave 5 (the mode treatments were only assigned after wave 4 fieldwork had been completed). For sample members with **Email** = 1, **Email wave** is a categorical variable that indicates the (most recent) wave at which an email address was supplied. **Partner's email** indicates whether an email address was known for

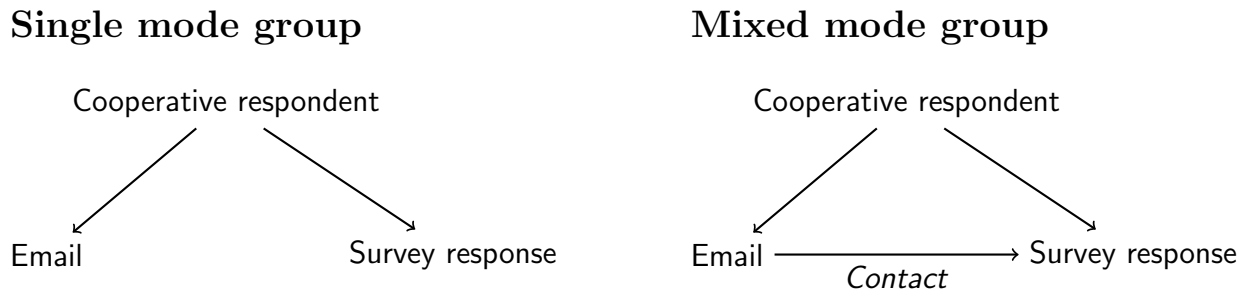
Figure 5.1: Wave at which respondents supplied an email address (percentages, N: 2916)



the sample member's partner. Fifteen additional variables are included in our models as controls for the selectivity effect in supplying email addresses. These include socio-demographic indicators such as age, gender, education and ethnicity, and a set of variables expected to be associated with propensity to respond in web mode. The latter set includes the presence of home broadband, regular internet use, and stated mode preference. All fifteen variables are described in the appendix. Three logistic regression models are developed:

Model 1 predicts participation based on the full sample. Here we exploit the random allocation into mode designs to test the interaction between **Email** and **Mode treatment**. This coefficient indicates whether the extra **Email** contact actually aids the response process. To understand why this is the case Figure 5.2 presents the expected relationships in the two randomized groups: single mode and mixed mode. If a relationship between **Email** and **Survey response** is found in the single mode design then this is due to a common cause, for example a general tendency to be cooperative. This is because people in the single mode were not contacted by email, so no direct causal effect of **Email** on **Survey response** is possible. On the

Figure 5.2: The link between the experimental data collection design and analysis strategy



other hand, if there is a difference in the effect of **Email** between single mode and mixed mode that can only be due to the effectiveness of email contact. This would mean that if a main effect of **Email** on participation is found, but no interaction with **Mode treatment**, then **Email** simply indicates a tendency to be cooperative, whereas an interaction in which the effect of **Email** on participation is stronger for the mixed mode group would suggest that email contact enhances response propensity. *Model 2* predicts survey participation conditional on being in the mixed mode treatment. This allows us to test the effect of **Email**, and interactions between this and other respondent characteristics, in the mixed mode context in which we are interested. Estimation of interactions will identify whether there are particular sample subgroups for whom the treatments are either effective or detrimental. *Model 3* predicts response mode conditional on participation, based on the mixed mode group alone. For parsimony, we include all fifteen control variables in each model, regardless of significance.

In each model, we perform two additional types of tests. We test interactions of **Email** and **Partner's email** with **Time in sample** as a test of whether any effect of email communications depends on time in sample. We also test to see whether **Email wave** is significant, as a test of whether effects depend on how long ago the email address was supplied (Figure 5.1 presents the proportion of people who provided an email address and, if they did, at which wave).

5.5 Results

a) Effect of contact by email on survey participation

Our main research question concerns the utility of additional contacts by email in the web-CAPI sequential mixed mode survey context. Results from *Model 1* show that the overall effect, in the entire sample, of obtaining the respondent's email address on participation is positive (OR 1.72, $P < 0.01$). This implies that those who provide an email address are more likely to participate in subsequent surveys. As mentioned previously this confounds both unobserved characteristics, such as general cooperativeness, with the direct effect of extra contact by email. To separate the two we must look at the interaction between **Email** and **Mode treatment** in *Model 1*. This is not significant (OR 0.7, $P > 0.1$), indicating no evidence that the effect on propensity to participate differs between the mixed-mode treatment (where the email address was used to make additional contacts) and the single mode CAPI treatment (where it was not used at all). Thus there is no evidence that using the respondent's email address for extra contacts is helpful in terms of gaining cooperation in a web-CAPI mixed mode survey. Moreover, the main effect of **Email** is not significant in *Model 2* (Table 5.2). This indicates no combined effect of obtaining an email address in previous waves and having the extra survey contact sent by email in a web-CAPI mixed-mode context.

b) Moderators of the effect of contact by email on survey participation

Though no overall effect of extra email contacts was found, it remains possible that effects may operate differentially across subgroups (for example, in opposite directions and hence cancelling out at the sample level). To test for such moderating effects we test interactions between each potential moderating variable and the randomly-allocated mode treatment. We find a significant interaction between **Mode treatment** and the indicator of whether the respondent's partner had provided an email address. In the mixed-mode context, those whose partners had pro-

Table 5.2: Odds ratios from logistic regression models of response and mode of response

Model	**1**	**2**	**3**
Dependent variable	**Response**	**Response**	**Response in web mode**
Analysis base	**Total sample**	**Mixed mode sample**	**Mixed mode respondents**
Mode treatment	0.75+		
Email	1.72**	1.17	1.77***
Mode treatment * Email	0.70		
Partner's email	0.79	1.63**	1.27
Mode treatment * Partner's email	2.01**		
Education			
A levels	0.85	0.84	1.44
GCSE or CSE	0.97	0.89	0.98
Vocational/none	0.77+	0.76	0.65*
Missing	0.74	0.33	0.29
Urban	1.12	1.36*	1.30
Female	1.13	1.10	1.02
Age	1.05**	1.03+	1.03
Age^2	1.00*	1.00	1.00
In couple	1.16	1.06	1.82**
White British	1.44*	1.40*	1.10
Employed	0.72**	0.79	0.86
Own house	1.43**	1.33*	2.43***
HH size	0.85***	0.84***	0.87*
Has mobile	1.26	1.61*	1.46
Broadband	1.73***	1.63**	3.60***
Daily internet	1.08	1.02	1.70**
Mode preference			
CATI	0.97	0.87	1.18
Postal	0.69*	0.79	1.48+
Web	0.61***	0.62**	1.88**
No preference	0.12***	0.14***	1.63+
Not by web[3]	1.32+	1.02	0.54**
Pseudo R-squared	0.18	0.16	0.24
N. of cases	2,522	1,665	1,142

Notes: For education the reference category is higher degree; for mode preference the reference category is CAPI;
$^{***}P < 0.001;^{**}\, 0.001 \leq P < 0.01;^{*}\, 0.01 \leq P < 0.05; +0.05 \leq P < 0.10$

vided an email address were significantly more likely to have participated, whereas in the single-mode CAPI context no such effect was observed. This is confirmed by a significant main effect of **Partner's email** in *Model 2* ($P < 0.01$) but not in *Model 1*. We find no interactions of **Email**, **Partner's email** or **Mode treatment** with any of the fifteen other indicator variables ($P > 0.05$) in *Model 1*. Thus, there is no evidence that any effect of **Email** or **Partner's email** acts differentially between sample subgroups or is moderated by whether the sample member has broadband internet access at home or whether they are a regular internet user. Furthermore, interactions with **Time in sample** were not significant, so there is no evidence that effects depend on time in sample. **Email wave** is not significant when substituted for **Email** in *Models 1* or *2*, so effects are not dependent on how recently the email address was supplied.

c) Effect of contact by email on mode of participation

Although extra contact by email does not appear to increase the overall propensity to participate in the survey it could still be beneficial if it increases the proportion of respondents who complete the survey by web instead of CAPI. With a sequential mixed-mode design, this would result in cost savings. We test such an effect in *Model 3*. Here we model the propensity to answer by Web as opposed to CAPI conditional on participating in the mixed mode survey. The significant main effect of **Email** indicates that sample members who had provided email addresses were more likely to respond in web mode rather than face-to-face (OR 1.77, $P < 0.01$). It should be noted that in this model we cannot take advantage of an experimental design, as there was no further randomisation to treatment (receiving additional contacts by email) within the mixed-mode group. Instead, we rely here on the inclusion of the other 15 independent variables to provide a control for differences in relevant respondent characteristics between those who provided an email address and those who did not. In so far as the controls are adequate, the result suggests that additionally providing the survey invitation and reminders by email increases the

propensity to respond by web rather than by face-to-face interview, thus reducing survey costs.

d) Moderators of the effect of contact by email on mode of participation

In extensions of *Model 3* (results not shown), we investigated interactions between **Email** and **Time in sample** and between **Email** and each of the 15 control variables. Two significant interactions were found: the effect of **Email** is stronger for those not in rural areas (**Email*Urban** $\hat{\beta} = 0.48$; P = 0.04) and for those who do not own their own house (**Email*Own** $\hat{\beta} = 0.30$; P = 0.01). Also, we note that the main effect of **Partner's email** is not significant in *Model 3*. This indicates that knowing the partner's email address (and hence being able to send survey invitations by email to the partner) is not associated with mode of participation in a mixed-mode survey, over and above the effect of knowing the respondent's email address (and hence being able to send survey invitations by email to the respondent). Note that this effect is net of the effect of having a partner and of having broadband at home (which may be associated with the probability of the partner having an email address), as indicators of both these characteristics are included as controls.

5.6 Discussion

It does not appear that obtaining a sample member's email address and using it to send a copy of the invitation letter and additional reminders in the first phase of a web-CAPI sequential mixed mode survey affects response propensity. However, email communication is associated with a higher propensity to respond in web mode as opposed to CAPI, an outcome that brings potential cost savings.

There are alternative explanations for the absence of an effect on participation propensity. Panel members may be relatively committed respondents and consequently less sensitive than others to influences on their participation propensity.

[3]Those that chose "Definitely would not" to the item: "And if next year we asked you to complete a questionnaire on the internet, how likely is it that you would complete the questionnaire?"

However, we doubt this explanation for two reasons. First, the proportion of persons issued to field at wave 5 who completed the individual interview was only 70.6% (see Table B.3), suggesting some scope for influence. Second, the absence of an interaction between **Email** and **Time in sample** implies that our results hold equally at the second and fifth annual waves of a survey. An alternative explanation may be that encountering URLs while offline and having to retain them until a suitable occasion when one is online, and entering passwords, may have become common and routine activities that are no longer a big barrier to participation (if they ever were). The extra convenience of being able to click a link may be rather trivial. Additionally, we do not know how many sample members actually received our emails. Some emails may have been diverted by spam filters (Fan and Yan, 2010) and others may simply have been left unopened. The email addresses provided by respondents may in some cases relate to accounts set up primarily for receipt of commercial mailings and the like. At wave 6, only 30% of our invitation emails were opened by the recipient (Wood and Kunz, 2014)[4].

Intriguingly, knowing the email address of the sample member's partner appears to increase response propensity in the mixed-mode context, but not in the single-mode CAPI context. This may indicate that making contact by email with both members of a couple has a positive effect (from the researcher's perspective) on both (recall that in most cases, the partner of a sample member will themselves be a sample member too in our design), whereas email contact with just one person has no effect on the response behaviour of that person.

With regard to the mode of participation in a sequential web-CAPI design, we find that sending invitation and reminder letters by email in addition to mail during the web phase increases the propensity of respondents to respond by web rather than CAPI (conditional on participation). This can certainly help to reduce survey costs. However, the effect is not observed in rural areas or amongst home owners. The identification of heterogeneous effects across socio-demographic groups such as

[4]13% bounced and 57% were unopened. For technical reasons we were unable to capture equivalent paradata at wave 5, the wave of the experiment reported here.

these might be useful for future research and for targeting purposes (Lynn, 2014b). Our findings regarding mode of participation require replication, preferably with an experimental allocation of email communications. We have tried to counter the possible selectivity in the process that leads to provision of an email address by controlling for relevant respondent characteristics (from socio-demographics to mode preference) and by interacting **Email** with all the controlling variables, but the possibility of unobserved heterogeneity remains.

In conclusion, the benefits of knowing the email address of sample members may be less than one might think. Researchers should evaluate carefully whether the intrusion and effort implied by a request to supply an email address are warranted. In a mixed mode context, as a means to improve participation, collecting email addresses may not be worthwhile. However, as a means to save costs by increasing the proportion of respondents who participate in web mode, the use of emails could be effective. Further research is required to replicate our findings in different populations, to better identify the determinants of mode of participation in sequential designs, and to learn more about the circumstances in which additional email contacts are worthwhile.

Chapter 6

Using equivalence testing to disentangle selection and measurement in mixed modes surveys

Abstract

Mixed modes are becoming increasingly popular in surveys. This approach can decrease costs and non-response bias. But in order to evaluate the utility of this approach we must separate selection and measurement effects of the different modes. In this paper I propose a new way of applying the front-door method to control for measurement differences between modes: equivalence testing with latent measurement models. A small simulation study will show how this approach works and how it can be biased if the assumptions of exhaustiveness and isolation are not true in the observed data.

6.1 Introduction

Using multiple modes of interview (i.e., face to face, telephone, web) to conduct surveys is increasingly popular as it can potentially lower costs while minimizing non-response bias (De Leeuw, 2005). But despite the increased popularity of mixed mode surveys there is still an acute need for methods to evaluate such designs. In order to gauge their effectiveness it is essential to separate the effects of modes on selection and measurement. Only then is it possible to investigate if the additional modes, usually more expensive, manage to include different types of individuals

and make the overall sample more representative. Additionally, identifying the measurement effects of the different modes can inform design decisions.

Most of the literature in mixed modes research has used multiple items to control for different selection propensity in modes in order to estimate measurement effects, also know as the *back-door method* (Pearl, 2009; Morgan and Winship, 2007; Vannieuwenhuyze et al., 2014). Recently, a different approach has been put forward, which aims to control for mode differences in measurement, known as the *front-door method* (Pearl, 1995, 2009; Morgan and Winship, 2007; Vannieuwenhuyze et al., 2014). This approach may prove an important development as situations can be envisaged where good back-door variables are not available but front-door ones are. Furthermore, considerable research and theory has been developed to estimate and explain measurement differences between modes. This knowledge can be fruitfully applied to the front-door method. While this approach has great potential, it hinges on the ability to find new variables that are able to control for measurement differences across modes.

The present paper will propose a new way to separate selection and measurement in mixed mode research by utilizing equivalence testing as a front-door method. While testing for equivalence has been previously used in the mixed mode literature (Révilla, 2013; Vannieuwenhuyze and Révilla, 2013; Klausch et al., 2013; Cernat, 2015a; Gordoni et al., 2011; Heerwegh and Loosveldt, 2011; Hox et al., 2015) it has been usually implemented to estimate measurement differences between modes after controlling for selection. The potential of this approach as a front-door method for estimating selection mode effects on a latent variable has been ignored so far. It is this point that this paper will elaborate on.

In order to show the potential of this method and it's assumptions the next two sections will present the main theoretical background of causal models and equivalence testing. Next, a simulation study will exemplify the method and the potential bias when assumptions do not hold. Finally, conclusions and limitations will be discussed.

6.2 Causal models and mixed modes

The fundamentals for the current discussion of causal analysis is based on the counterfactual model which stipulates the existence of multiple causal states to which the population of interest could be exposed. In the simple case of a mixed mode design with two modes each individual could answer either in the first mode, m_1, or in the second one, m_2. Using the notation of Vannieuwenhuyze et al. (2014) this will be denoted by D and is called *mode of data collection*. Nevertheless, in a survey each respondent participates only using one mode, the *mode group*, denoted by G_δ (where δ stands for the design used). Figure 6.1 graphically presents this situation. In the ideal counterfactual data we would have both D and G_δ and they would not be related (situation a). Unfortunately, most of the real data has only one observation per individual and thus the two variables can't be distinguished (situation b).

Usually, the interest lies with the mean of a variable in the reference mode: $\mu_{m_1} = E(Y|D = m_1)$. Nevertheless, calculating this is not possible with observed data as it requires counterfactual information:

$$\mu_{m_1} = \mu_{m_1 m_1}\tau_{m_1} + \mu_{m_1 m_2}\tau_{m_2} \tag{6.1}$$

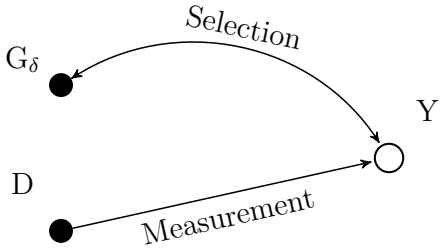
where μ_{dg} is the conditional mean $E(Y|D = d, G_\delta = g)$. In this equation $\mu_{m_1 m_1}$ can be observed in the data as the people that answered using m_1 while $\mu_{m_1 m_2}$ is a counterfactual as it represents what would the respondents from m_2 would have answered had they participated in m_1. Here τ_g represents the propensity to answer in each group: $P(G_\delta = g)$.

Using this notation we can estimate the selection and the measurement effects:

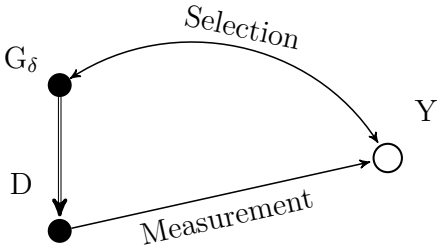
$$S_{m_1}(\mu) = \mu_{m_1 m_1} - \mu_{m_1 m_2} \tag{6.2}$$

Figure 6.1: Counterfactual models for separating selection and measurement in a mixed mode design. Adapted from Vannieuwenhuyze et al. (2014).

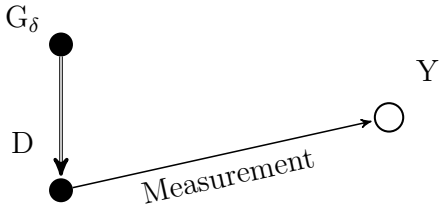
a. Counterfactual data



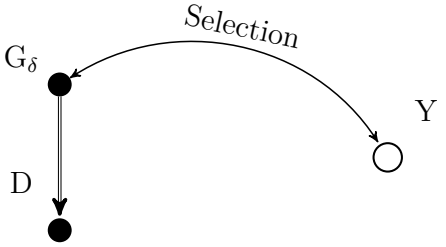
b. Mixed mode data



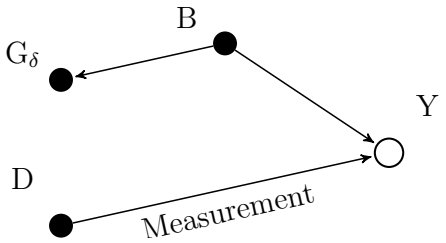
c. No selection



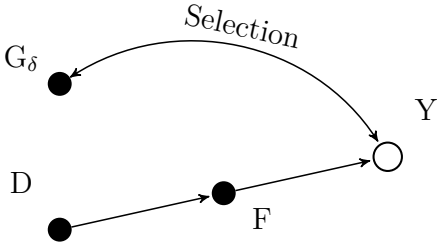
d. No measurement



e. Back door



f. Front door



$$M_{m_1}(\mu) = \mu_{m_2 m_2} - \mu_{m_1 m_2} \tag{6.3}$$

The selection effect, $S_{m_1}(\mu)$, would be different from zero only if the people in the two modes would have different different means had they all answered in m_1. Similarly, the measurement effect, $M_{m_1}(\mu)$, is given by the difference between the respondents in m_2 and those in m_1 if they had answered in the second mode. These formulae highlight the importance of estimating the counterfactual in separating selection and measurement, this being essential for the evaluation of mixed mode designs.

Using only the observed data does not enable the estimation of the two types of mode effects. As a result, a series of models have been put forward in order to estimate the counterfactuals. The causal literature has presented three main techniques: instrumental variables, the back-door approach and the front-door approach (Pearl, 2009; Morgan and Winship, 2007; Vannieuwenhuyze et al., 2014). The focus here will be on the latter two.

The *back-door method* aims to use a series of covariates (B in Figure 6.1e) that explain both the variable of interest, Y, and the survey mode (G_δ). It has been shown that by controlling for such variables it will be possible to calculate the counterfactual $\mu_{m_1 m_2}$ and, thus, calculate the measurement effect (Pearl, 2009; Morgan and Winship, 2007; Vannieuwenhuyze et al., 2014).

While this technique has been used repeatedly in the mixed mode field it does have two important assumptions. The first one is the *ignorable mode selection assumption*. This implies that the B variables will capture the entire relationship between mode and the variable of interest Y (i.e., selection effect into survey mode). When this assumption does not hold the estimates of selection and measurement effects of mode on Y will be biased as they will still be confounded with selection on unmeasured B variables. The second assumption is the *mode insensitivity assumption*. This means that there is no relationship between B and D. In practice this implies that the measurement of the controlling variables is not influenced by the mode of measurement.

The back door has been applied in the mixed mode literature multiple times using techniques such as regression (e.g. Jäckle et al., 2010), matching (e.g. Lugtig et al., 2011), weighting (e.g. Hox et al., 2015) and controlling for covariates in Structural Equation Modelling (e.g., Heerwegh and Loosveldt, 2011).

Another approach to separating selection and measurement in mixed mode designs is the *front-door method* (Pearl, 1995, 2009; Morgan and Winship, 2007; Vannieuwenhuyze et al., 2014). Here the aim is to find a set of variables F (Figure 6.1f) that explain the measurement effect of the mode on the variable of interest.

As with the previous approach the front-door also makes a number of assumptions. The first one is the *exhaustiveness assumption*. This implies that the F-variables capture the entire causal effect of D on Y. If this is not true, part of the estimated selection differences will include differential measurement. Then, the *isolation assumption* requires that F is independent of G_δ; if it does not hold, then F will also include part of the selection effect.

The front-door approach is relatively new in the causal literature and has been rarely used in the mixed mode field (Vannieuwenhuyze et al., 2014). Although the assumptions of the method are similar to those of the back-door the variables used in the two procedures to separate selection and measurement are very different. Increasing the use of the front-door will hinge on finding appropriate variables to control for measurement differences. Raising awareness of this procedure and developing new ways to implement it in the field of mixed modes will provide researchers with new tools to evaluate surveys that combine multiple modes. Next, we turn to latent models and how they can be used to estimate and correct for relative bias in measurement.

6.3 Equivalence testing and measurement

The use of latent variables in psychology, sociology or education has developed considerably in the last half a century in an aim to control for the inevitable fallibility of observed items and in order to get closer to substantial concepts used in theory.

This development has been based on the Classical Test Theory (Lord and Novick, 1968) and has been extended with the use of latent variables in Structural Equation Modeling, Latent Class and Item Response Theory. These approaches assume that there is an underlying, unobserved, concept of interest that is measured with error by observed variables.

One such general model is the Confirmatory Factor Analysis (Bollen, 1989). Here we assume that a vector p of observed items, y, are explained by an m set of underlying continuous latent variables, ξ:

$$y^{(g)} = \upsilon^{(g)} + \Lambda^{(g)}\xi^{(g)} + \epsilon^{(g)} \tag{6.4}$$

where Λ is a $p * m$ matrix of factor loadings, υ is a vector of intercepts or thresholds and ϵ is a p vector of residuals (variances) independent of ξ and with a mean of zero (Bollen, 1989). The superscript g indicates that the coefficients may vary across g groups. Let μ_ξ and ϕ_ξ be the mean and the variance of ξ.

In this framework the loadings, Λ, and residuals, ϵ, can be considered to estimate the reliability of the items (Bollen, 1989). The intercept, or the threshold when the observed variables are categorical, is linked to the systematic part of the model. Variations of these quantities are also know in the Item Response Theory as discrimination and difficulty.

This measurement approach has been further extended to estimate relative bias by comparing these models across groups (Millsap, 2012; Meredith, 1993; Steenkamp and Baumgartner, 1998). Because researchers are usually interested in ξ it is essential that this is measured similarly (i.e., be equivalent or invariant) in each group of interest. If this is not the case, then any use of the latent variable may confound differences in measurement with substantive differences.

In order to evaluate whether the measurement model is equivalent across groups, and relative measurement error is the same, a series of nested models are tested. In each group different levels of equality restrictions are added across groups. Usually, the procedure starts with a general model, called the *configural model*, which

assumes that the same structure is found across groups, but no equality constraints are imposed on the coefficients. If this model is found to fit the data, then a set of restrictions can be imposed on the Λ coefficients. If this model also holds (i.e., if it's not significantly worse than the configural model) the model is considered *metric equivalent* across groups (Steenkamp and Baumgartner, 1998). Next, a new set of restrictions can be added on the intercepts/thresholds, υ. If this model is accepted (i.e., fits the data well) then it is considered *scalar equivalent* (Steenkamp and Baumgartner, 1998), *strong factorial equivalent* (Meredith, 1993) or *first-order measurement invariant* (Millsap, 2012). The model can further be restricted to *strict factorial invariance* (Meredith, 1993) or *second-order invariance* (Millsap, 2012) by imposing equal random errors, ϵ. It should be noted that in order to compare means of the latent variable(s), μ_ξ, scalar equivalence needs to be found while in order to compare variances, ϕ_ξ, strict factorial invariance must be accepted.

The different levels of cross-group equality presented above are relatively strict and are hard to find in real-life data. As such, the concept of *partial equivalence* has been put forward (Byrne et al., 1989; Steenkamp and Baumgartner, 1998). This implies that even if not all the coefficients are equal across groups unbiased coefficients of ξ can be estimated if at least two items are equivalent and if the differences found on the other items are controlled for. This compromise has been found valuable as real world data has shown this to be quite common (e.g., Davidov, 2008).

While equivalence testing has become very popular due to the methodological and substantive insights it brings it nevertheless has a number of limitations. One of them refers to the fact that it can be implemented only when multiple items (preferably more than two for each ξ) of the same dimension are measured (Alwin, 2007). Secondly, the procedure estimates only relative bias. The measurement model may be the same across groups but may lack validity. Thirdly, the usual procedure for ascertaining the level of equivalence is exploratory and may capitalize on chance. Finally, the procedure cannot deal with certain types of systematic errors. For example, if primacy (i.e., tendency of selecting the first category irre-

gardless of the question) is higher in all the items of one group then the difference will be included in the mean of ξ, thus confounding substantive and measurement differences across groups. This can be ameliorated by including the systematic errors in the model as has been done with acquiescence (Billiet and Davidov, 2008; Billiet and McClendon, 2000), method (Campbell and Fiske, 1959; Andrews, 1984; Saris and Gallhofer, 2007a) or extreme response style (Kankaraš et al., 2011).

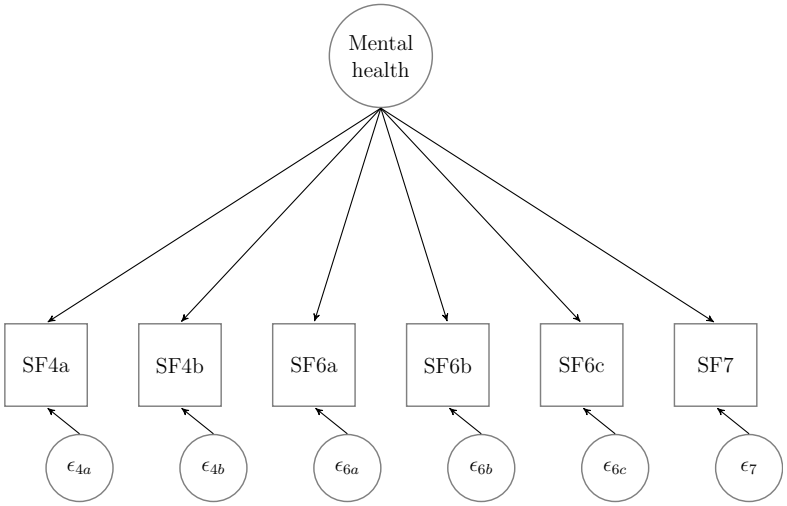
As mentioned at the beginning of the article, the equivalence testing procedure has been used a number of times in the mixed mode literature. The most typical use is to estimate measurement differences between modes after controlling for selection using a set of back-door variables, usually socio-demographics items (Hox et al., 2015; Klausch et al., 2013; Gordoni et al., 2011). Alternatively, it has been used to compare measurement differences of mode designs (Biemer, 2001) when these were randomly allocated (e.g., Cernat, 2015a). Previous research has also considered one of the limitations presented previously and have included other systematic errors in the model when comparing modes, such as acquiescence (Heerwegh and Loosveldt, 2011) or method (Révilla, 2013).

6.4 Equivalence testing as front-door approach

Given the the discussion so far, a natural question arises: is it possible to use equivalence testing, which was developed to estimate and control for differences in measurement, as a front-door to separate mode effects on selection and measurement? Because we do not expect mode to have a causal impact on the latent variable any differences found on this dimension can be due to selection, measurement or a combination of the two. Using equivalence testing we should be able to control for measurement differences, if these appear in the form of partial equivalence.

To see if this is the case and understand how results may be biased if assumptions don't hold a small simulation study will be presented below. Let's assume we want to measure mental health and we want to know whether people with different levels of health select into modes. One possible way to measure this is with items from the

Figure 6.2: Measurement model to be tested for equivalence.



SF12 scale (Ware et al., 2007). SF12 is a scale developed to measure both physical and mental health. As such, we will choose only those items that measure the latter sub-dimension (Figure 6.2).

In order to have plausible values for the population model we will use results from the Understanding Society Innovation Panel wave 5 (Cernat, 2015a; McFall et al., 2013). Applying the model in Figure 6.2 to these data we retrieve the following values that will be used as the true/population scores in the simulation study for the first group, m_1 (we will call these *Coef. 1*):

$$y = \upsilon + \Lambda\xi + \epsilon \tag{6.5}$$

$$\begin{bmatrix} SF_{4a} \\ SF_{4b} \\ SF_{6a} \\ SF_{6b} \\ SF_{6c} \\ SF_{7} \end{bmatrix} = \begin{bmatrix} 4.4 \\ 4.6 \\ 2.5 \\ 2.8 \\ 4 \\ 4 \end{bmatrix} + \begin{bmatrix} 1 \\ 0.9 \\ -0.45 \\ 0.5 \\ 0.7 \\ 0.8 \end{bmatrix} \begin{bmatrix} \xi \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0.2 \\ 0.6 \\ 0.7 \\ 0.6 \\ 0.6 \end{bmatrix}$$

Let us further assume that the mean and variance for the mode of interest are $\mu_\xi^{m1} = 0$ and $\phi_\xi^{m1} = 1$. Furthermore, selection effects on the latent variable for the second mode will be added: $\mu_\xi^{m2} = 1.5$ and $\phi_\xi^{m2} = 1.5$. We know from the literature

on equivalence testing that estimating a Multi-Group Confirmatory Factor Analysis assuming strict factorial invariance when only selection on the latent variable is present will lead to unbiased estimates (Hox et al., 2015; Meredith, 1964). Now lets assume that the second mode also has a measurement effect. This can be included in the model by imposing different intercepts, loadings and random errors in m_2 (which we will call *Coef. 2*):

$$y = \upsilon + \Lambda\xi + \epsilon \tag{6.6}$$

$$\begin{bmatrix} SF_{4a} \\ SF_{4b} \\ SF_{6a} \\ SF_{6b} \\ SF_{6c} \\ SF_{7} \end{bmatrix} = \begin{bmatrix} 5 \\ 4.6 \\ 2.5 \\ 2.8 \\ 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 1 \\ 0.5 \\ -0.85 \\ 0.5 \\ 0.7 \\ 0.8 \end{bmatrix} \begin{bmatrix} \xi \end{bmatrix} + \begin{bmatrix} 0.4 \\ 0.2 \\ 0.6 \\ 0.95 \\ 0.6 \\ 0.6 \end{bmatrix}$$

We expect that ignoring the confounding of selection and measurement in the two modes would lead to the biased estimation of the former. This can be clearly seen in the case 1 of Table 6.1 [1]. The mean and the variance of the selection in the second group is biased when we ignore measurement differences: a bias of 31 for the mean and 20 for the variance for the selection on the latent variable in the second mode. From the previous section we expect that if we are able to find partial equivalence between the modes then we can control for differences in measurement and estimate unbiased mode selection effects on the latent variable. By calculating the same model but freeing the coefficients that are different in the two groups we estimate the correct values for the mode selection effects (case 2 in Table 6.1). This exemplifies how partial equivalence testing can be used as a front door method for estimating selection on a latent variable of interest.

While this is very encouraging we also know that this approach has two impor-

[1] The simulations have been run in Mplus 7.2. Sizes of 2000 respondents were assumed for each group. A 1000 repetitions were used.The syntax can be found as supplemental files.

tant assumptions. The first one, exhaustiveness, implies that the partial equivalence captures all the measurement differences between the two modes. If this is not true then the selection effect will be biased. To test this let us imagine that in addition to the selection and measurement differences already included in the model, there is also a type of systematic error in the second mode. This can take different forms such as acquiescence, social desirability, extreme response styles or recency/primacy. Here we will assume that acquiescence or primacy increases the chances of choosing the first category in the second mode. This is implemented in the model by adding a latent variable in the second group. This has loading of 1 on all the observed variables and a mean and variance of 1 (Billiet and Davidov, 2008; Billiet and McClendon, 2000). As expected, if this type of mode difference in measurement is ignored, then the estimate of selection will be biased (case 3 in Table 6.1): Means Square Error for $\mu_{\xi}^{m_2}$ and for $\phi_{\xi}^{m_2}$ are approximately 1. If appropriate measures are in the data, for example if balanced items are used for the items, then the response style can be modeled. When this is included (case 4 in Table 6.1) selection effects will not be biased. This highlights both a limitation of the model but also it's flexibility in including multiple types of systematic errors.

A second assumption of the front-door method is isolation. This implies that there are no other unobserved variables that have an impact both on measurement and selection. We can think of multiple theoretical situations when this may not be plausible. For example, people with lower working memory may have more measurement error in an auditory mode than a visual one and may also auto-select in one of them. To model such a situation let us imagine we have four groups: the reference mode with high working memory (m_{1a}) and with low working memory (m_{1b}), and the second mode with high working memory (m_{2a}) and with low working memory (m_{2b}). If isolation is not true in the population then working memory will have a differential effect on measurement and selection in the two modes. We can model this by imposing *Coef. 1* in the first three groups: m_{1a}, m_{1b} and m_{2a}. To estimate measurement differences for the fourth group we will impose *Coef.2* on m_{2b}.

Table 6.1: Simulation results

Nr.	Model: Population	Model: Estimation	Coefficient	Population	Model	Bias*	M.S.E**
1	Selection + partial equivalence	Selection	$\mu_\xi^{m_2}$	1.5	1.97	31.33	0.22
			$\phi_\xi^{m_2}$	1.5	1.8	20.00	0.09
2	Selection + partial equivalence	Selection + partial equivalence	$\mu_\xi^{m_2}$	1.5	1.5	0.00	0.00
			$\phi_\xi^{m_2}$	1.5	1.5	0.00	0.00
3	Selection + partial equivalence + response style	Selection + partial equivalence	$\mu_\xi^{m_2}$	1.5	2.48	65.33	0.96
			$\phi_\xi^{m_2}$	1.5	2.56	70.67	1.14
4	Selection + partial equivalence + response style	Selection + partial equivalence + response style	$\mu_\xi^{m_2}$	1.5	1.5	0.00	0.00
			$\phi_\xi^{m_2}$	1.5	1.5	0.00	0.00
5	Selection + partial equivalence + non-isolation	Selection + partial equivalence	$\mu_\xi^{m_{2a}}$	1	1.75	75.00	0.07
			$\phi_\xi^{m_{2a}}$	1	2.14	114.00	0.02
			$\mu_\xi^{m_{2b}}$	2	1.75	-12.50	0.07
			$\phi_\xi^{m_{2b}}$	2	2.14	7.00	0.02
6	Selection + partial equivalence + non-isolation	Selection + partial equivalence + non-isolation	$\mu_\xi^{m_{2a}}$	1	1	0.00	0.00
			$\phi_\xi^{m_{2a}}$	1	1	0.00	0.00
			$\mu_\xi^{m_{2b}}$	2	2	0.00	0.00
			$\phi_\xi^{m_{2b}}$	2	2	0.00	0.00

** Bias = 100*(Population - Sample)/Population; ** Mean Square Error = variance of sample estimation + Bias^2.*

To simulate different selection we will impose the same mean and variance for the first mode $\mu_{\xi}^{m_{1a}} = \mu_{\xi}^{m_{1b}} = 0$ and $\phi_{\xi}^{m_{1a}} = \phi_{\xi}^{m_{1b}} = 1$ and differential selection within mode 2: $\mu_{\xi}^{m_{2a}} = 1$, $\phi_{\xi}^{m_{2a}} = 1$, $\mu_{\xi}^{m_{2b}} = 2$ and $\phi_{\xi}^{m_{2b}} = 2$.

In the real data, if we do not measure working memory then we assume that everything within each mode is equal (i.e., coefficients of $m_{1a} = m_{1b}$ and $m_{2a} = m_{2b}$). The theoretical expectation is that this indeed will bias the estimate of selection in the latent variable. This is obvious in case 5 of Table 6.1, where the coefficients for selection in the two subgroups of the second mode are equal but both coefficients have systematic error with bias ranging from 7 for $\phi_{\xi}^{m_{2b}}$ to 114 for $\phi_{\xi}^{m_{2a}}$. The last case of the simulation study shows once again that this assumption can be freed if we measure working memory in the data and if we include it in our model. The estimation of selection on the latent variable is unbiased and the model controls for differential measurement and selection.

6.5 Conclusions and discussion

This paper has shown how it is possible to conceptualize equivalence testing as a front-door method to estimate selection on a latent variable. While this technique has been used multiple times in the field of mixed modes it has yet to be considered on its own terms as a method to deal with the confounding of selection and measurement. The simulation study has shown that the method will work and give unbiased estimates.

That being said, the model does make two important assumption: isolation and exhaustiveness. The simulation has shown that indeed when these do not hold in the population the sample estimates of selection will be biased. Nevertheless, the method is flexible enough to give users the opportunity to include any potential biasing factors as covariates. This makes for a very versatile method for disentangling selection and measurement.

Equivalence testing has it's own limitations as a statistical method, such as the need for multiple items or capitalization on chance. This may lead to other types

of biases when the method is applied to the real world data. The paper has not tackled this issue directly but there is considerable ongoing research that should reduce these issues in the future (e.g., Asparouhov and Muthén, 2014).

The paper has only highlighted the utility of the approach and possible limitations. In order to make it more attractive for real world applications further research is needed. For example, a thorough study that simulates multiple types of models with varying degrees of miss-specification (e.g., multiple types of errors, multiple types of unobserved covariates) may indicate to users the degree of bias they can expect when applying this method. Similarly, developing methods to utilize the information estimated using this approach for other purposes, such as creating weights or correcting substantive models, should be pursued.

6.6 Advice for practitioners

In this section I will highlight a procedure that practitioners can implement in their own analyses using the method presented in this chapter. This follows three distinct steps. It should be noted that this is just one possible approach, and practitioners should adapt this as it best fits their own needs.

The first step is to estimate the total mode differences on the latent variable. This can be done by implementing a multi group CFA (or latent class, depending on the variables used). In this model all loadings and intercept/thresholds should be assumed equal but the mean and the variance of the latent variable should be freely estimated. The differences between groups on the latent variable would represent the total mode effect.

The second step is to use fit indicators to free loadings and intercept/thresholds that are significantly different between the groups/modes. These represent measurement differences and freeing them would enable a correct estimation of the selection effect on the latent variable. If partial invariance is found and the two assumptions, exhaustiveness and isolation, hold any difference between the two latent variables represent mode selection effects.

Lastly, the researcher should investigate to what degree the two assumptions hold. This can be done by including in the model variables that can explain differences in measurement between modes such as: if someone else was present during the interview (linked to social desirability), time latencies (linked to satisficing), age or education (linked to cognitive ability). Including such variables with effects on the observed items of the scale would help control for other causes in measurement differences across modes and make the exhaustiveness assumption more plausible. Additionally, sensitivity analyses should be done as presented in the missing not at random literature. This could be easily implemented in the latent variable procedure. For example, the researcher could postulate a method effect as a latent variable that has different effects in the mode groups. Then the impact on the estimate of mode selection can be investigated.

Chapter 7

Conclusions

That concludes our explorations of the impact of mixed modes in longitudinal studies. During the last six chapters we have explored eight waves of data from two large scales surveys on two continents. Furthermore, a simulation was used to propose a new method to deal with separating selection and measurement mode effects. In this process I hope that the research presented here has contributed to the survey methodology literature in new and innovative ways.

Contribution to the literature

Theoretical contributions

As of yet there has been only limited research on the topic of mixed modes in longitudinal studies. As this is becoming more relevant for survey practitioners and users there is a need to understand the characteristics that make longitudinal studies distinct and how these might interact with mixed mode designs. To the knowledge of the author Chapter 2 is the first one to highlight the main characteristics of panel data that can interact to mixed modes designs (with the notable exception of Dillman, 2009) and to discuss how and when in the panel studies the effects can appear. The framework is still in its incipient form and can be improved in the future but it is hoped to give a theoretical starting point for research on this topic.

Similarly, the PhD has contributed to the theoretical understanding of mixed modes research by integrating two techniques that although have been applied previously they have never been used together before. Recently Vannieuwenhuyze et al.

(2014), inspired by the causal literature (Pearl, 1995, 2009; Morgan and Winship, 2007), has proposed the use of the front-door method in mixed mode research. This proposes that instead of trying to model the characteristics of people that auto-select in a certain mode (the typical approach in the literature, called the back-door approach) we can also control for possible measurement mode differences in order to estimate selection effects. This new perspective would enable us to utilize the knowledge regarding measurement mode effects accumulated in the literature as a way to disentangle it from selection. In Chapter 6 I propose to use equivalence testing as a front door approach. This is a statistical technique that uses models such as Confirmatory Factor Analysis in order to estimate and correct for measurement differences across groups. Although it has been used in the field previously (e.g., Hox et al., 2015) it has never been conceptualised in this way before. It is hoped that this new approach will become in time a useful tool in the arsenal of survey methodologists and users in the attempt to understand mode effects in non-experimental data.

Empirical contribution

Research carried out throughout this PhD has tackled different aspects of the mixed mode design with a special focus on panel data. The analyses carried aim to support other survey methodologists and users in collecting and utilizing this kind of data. Here I summarise the main empirical findings from this research:

- There are small differences differences in random error between a face-to-face single mode and a sequential telephone - face-to-face mixed mode design as 32 out of 33 items have the same reliability (Chapter 2).

- There are very small differences in the way mental and physical health (as estimated by the SF12) are measured in a single mode (face-to-face) and a mixed mode (sequential telephone - face-to-face), both in the wave in which two designs was implemented as well as in subsequent waves (Chapter 3).

- The mixed mode design overestimates change of individuals substantially for

four out of 12 items of the SF12 (Chapter 3).

- The measurement of depression, physical activity and religiosity is the same between face-to-face and telephone modes. Nevertheless, these two modes measure depression and physical activity systematically different from the way they are measured in the Web survey (Chapter 4).

- Extra invitations by email does not increase propensity to participate in a Web - face-to-face sequential mixed mode design (Chapter 5).

- Extra invitations by email may increase participation in Web as opposed to face-to-face, thus potentially saving costs for survey agencies (Chapter 5).

- Equivalence testing does indeed work as a front-door method as long as the two main assumptions, exhaustiveness and isolation, hold (Chapter 6).

An overall findings of the research is that mode differences can be small or absent when the modes are similar and when the design decisions were made to minimize these differences (Chapters 2, 3 and 4). One example is the Health and Retirement Study that stopped using showcards when they introduced the telephone component in order to minimize mode effects. That being said, differences between mode designs and modes can still be present as exemplified by the differences in estimates of change in Chapter 3 and between interviewer modes and the Web answers when measuring depression in Chapter 4.

Advice to survey practitioners

It is hoped that the research presented previously not only contributes to the academic literature but it can also inform decision making for survey managers. As such, the research leads to a number of practical recommendations. Firstly, when mixing multiple modes in a survey **consider combining modes that are similar**. For example, the Health and Retirement Study combines face-to-face and telephone without showcards. This has been shown to have very small differences (see Chapter 4 for an example) and, as such, assumptions about equivalent measurement and

selection across modes is more plausible. Similarly, a design such as the used by UKHLS that implements a self-completion section as part of the face-to-face survey for the items that are more sensitive and more prone to mode effects would facilitate combinations with a Web mode.

Secondly, when collecting data using a mixed mode design **include an experimental group in the main single mode approach**, if possible. As has been shown in Chapters 2 - 5 this facilitates the mode comparison and can inform decisions both about the future design of the survey and about ways in which to correct for mode effects that could bias substantial analyses.

Thirdly, if you are designing a Web - face-to-face sequential survey **consider the trade-off between the costs of collecting emails and the potential benefits**. In Chapter 5 we have shown that the extra contact by email does not lead to increased response rates. Nevertheless, there are indications that the extra contact by email might increase the propensity to answer by Web instead of face-to-face, thus leading to potential cost savings.

Lastly, when multi-item scales are available in a mixed mode survey **consider using equivalence testing as a way to control for measurement differences**. In addition to the benefit of controlling for measurement mode effects at no extra costs it is also a flexible framework in which additional information about measurement and selection mode effects can be included in the model.

Future research directions

As mentioned before this is only part of a growing literature on mixed mode designs. In order to continue this development I propose two distinct directions for future research.

Firstly, it is important that research in this field moves beyond indicators of data quality such as number of "Don't Know" answers and reliability in order to investigate coefficients that are more important for substantial research. In this PhD this was done by comparing estimates of change across mode designs. This

can be further developed to substantive models such as regression coefficients, survival analysis, etc. This would inform more the users of mixed mode data on the potential pitfalls when ignoring these methodological issues. Additionally, it would also provide survey methodologists with what might be a more relevant metric for the impact of mode design.

Secondly, an area of future research can be the development of the front door approach to separating selection and measurement mode effects. It has been shown in Chapter 6 the it is possible to use some of the measurement models regularly used in survey methodology, such as Multi-Group Confirmatory Factor Analysis, as a front door method. Although this is not perfect it is a very flexible framework where additional information both for measurement and selection mode effects can be easily included in order to make the assumptions of the model more plausible. A different strand of research can also investigate how this new approach compares to the back-door typically used in the literature in order to inform practitioners on best practices and advantages/disadvantages in using each approach.

Bibliography

Alwin, D. F. (2007). *The margins of error: a study of reliability in survey measurement*. Wiley-Blackwell.

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

Aquilino, W. S. (1992). Telephone versus face-to-face interviewing for household drug use surveys. *Substance Use & Misuse*, 27(1):71–91.

Aquilino, W. S. (1998). Effects of interview mode on measuring depression in younger adults. *Journal of Official Statistics*, 14(1):15–29.

Asparouhov, T. and Muthén, B. (2010). Weighted least squares estimation with missing data. *Technical Report*, pages 1–10.

Asparouhov, T. and Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4):495–508.

Bandilla, W., Couper, M. P., and Kaczmirek, L. (2012). The mode of invitation for web surveys. *Survey Practice*, 5(3).

Bandilla, W., Couper, M. P., and Kaczmirek, L. (2014). The effectiveness of mailed invitations for web surveys and the representativeness of Mixed-Mode versus internet-only samples. *Survey Practice*, 7(4).

Béland, Y. and St-Pierre, M. (2008). Mode effects in the canadian community health survey: A comparison of CATI and CAPI. In Lepkowski, J. M., Tucker, C., Brick,

M., De Leeuw, E., Japec, L., Lavrakas, P., Link, M., and Sangster, R., editors, *Advances in telephone survey methodology*, pages 297–314. John Wiley & Sons, New York.

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.

Biemer, P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics*, 17(2):295–320.

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.

Bishop, G. and Smith, A. (2001). Response-Order effects and the early gallup Split-Ballots. *Public Opinion Quarterly*, 65(4):479–505.

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

Bollen, K. and Curran, P. (2005). *Latent Curve Models: A Structural Equation Perspective*. Wiley-Interscience, 1 edition.

Bosnjak, M., Neubarth, W., Couper, M. P., Bandilla, W., and Kaczmirek, L. (2008). Prenotification in Web-Based access panel surveys: The influence of mobile text messaging versus E-Mail on response rates and sample composition. *Social Science Computer Review*, 26(2):213–223.

Bradburn, N. M., Sudman, S., Blair, E., and Stocking, C. (1978). Question threat and response bias. *Public Opinion Quarterly*, 42(2):221–234.

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

Buck, N. and McFall, S. (2012). Understanding society: design overview. *Longitudinal and Life Course Studies*, 3(1):5–17.

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. (2012). Understanding society innovation panel wave 4: Results from methodological experiments. Working Paper 2012-06, University of Essex, ISER, Colchester.

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.

Byrne, B., Shavelson, R., and Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105(3):456.

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.

Cannell, C., Groves, R., Magilavy, L., Mathiowetz, N., and Miller, P. (1987). An experimental comparison of telephone and personal health surveys. Technical Series 2 106, National Center for Health Statistics.

Cernat, A. (2015a). Impact of mixed modes on measurement errors and estimates of change in panel data. *Survey Research Methods*, 9(2):83–99.

Cernat, A. (2015b). The impact of mixing modes on reliability in longitudinal studies. *Sociological Methods & Research*, 44(3):427–457.

Cernat, A. (2015c). Using equivalence testing to disentangle selection and measurement in mixed modes surveys. *Understanding Society Working Paper Series*, (01):1–13.

Cernin, P. A., Cresci, K., Jankowski, T. B., and Lichtenberg, P. A. (2010). Reliability and validity testing of the Short-Form health survey in a sample of Community-Dwelling african american older adults. *Journal of Nursing Measurement*, 18(1):49–59.

Chan, K. S., Orlando, M., Ghosh-Dastidar, B., Duan, N., and Sherbourne, C. D. (2004). The interview mode effect on the center for epidemiological studies depression (CES-D) scale: an item response theory analysis. *Medical care*, 42(3):281–289.

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.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? the impact of making inappropriate comparisons in cross-cultural research. *Journal of personality and social psychology*, 95(5):1005–1018. PMID: 18954190.

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

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.

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.

Couper, M. and Ofstedal, M. B. (2009). Keeping in contact with mobile sample members. In Lynn, P., editor, *Methodology of longitudinal surveys*, pages 183–203. Wiley, Chichester.

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

Couper, M. P. (2013). Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods*, 7(3):145–156.

Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the european social survey. *Survey Research Methods*, 2(1):33–46.

Davidov, E., Meuleman, B., Billiet, J., and Schmidt, P. (2008). Values and support for immigration: A Cross-Country comparison. *European Sociological Review*, 24(5):583–599.

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.

De Leeuw, E. D. and van der Zouwen, J. (1988). Data Quality in Telephone and Face to Face Surveys: A Comparative Meta-Analysis. In Groves, R., Biemer, P.,

Lyberg, L., Massey, J., Nicholls II, W., and Waksberg, J., editors, *Telephone Survey Methodology*, Wiley Series in Probability and Mathematical Statistics, pages 283–299. John Wiley & Sons, New York.

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

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. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods*, 17(1):30–52.

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

Duffy, B., Smith, K., Terhanian, G., and Bremer, J. (2005). Comparing data from online and face-to-face surveys. *International Journal of Market Research*, 47(6):615–639.

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

Fan, W. and Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior*, 26(2):132–139.

Fricker, S., Galesic, M., Tourangeau, R., and Yan, T. (2005). An experimental

comparison of web and telephone surveys. *Public Opinion Quarterly*, 69(3):370–392.

Gmel, G. (2000). The effect of mode of data collection and of non-response on reported alcohol consumption: a split-sample study in switzerland. *Addiction*, 95(1):123–134.

Gordoni, G., Schmidt, P., and Gordoni, Y. (2011). Measurement invariance across Face-to-Face and telephone modes: The case of Minority-Status collectivistic-oriented groups*. *International Journal of Public Opinion Research*.

Greenfield, T. K., Midanik, L. T., and Rogers, J. D. (2000). Effects of telephone versus face-to-face interview modes on reports of alcohol consumption. *Addiction*, 95(2):277–284.

Groves, R. (1979). Actors and questions in telephone and personal interview surveys. *Public Opinion Quarterly*, 43(2):190–205.

Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls II, W., and Waksberg, J., editors (1988). *Telephone Survey Methodology*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., and Tourangeau, R. (2008). *Survey Methodology*. Wiley-Blackwell, 2nd edition edition.

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

Groves, R. M. (1990). Theories and methods of telephone surveys. *Annual Review of Sociology*, 16(1):221–240.

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

Groves, R. M. and Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. Wiley-Interscience, New York, 1 edition edition.

Groves, R. M., Singer, E., and Corning, A. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *The Public Opinion Quarterly*, 64(3):299–308.

Hays, R. D., Kim, S., Spritzer, K. L., Kaplan, R. M., Tally, S., Feeny, D., Liu, H., and Fryback, D. G. (2009). Effects of mode and order of administration on generic Health-Related quality of life scores. *Value in Health*, 12(6):1035–1039.

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. and 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(1):49–63.

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

Herzog, A. R., Rodgers, W. L., and Kulka, R. A. (1983). Interviewing older adults: A comparison of telephone and Face-to-Face modalities. *The Public Opinion Quarterly*, 47(3):405–418. ArticleType: research-article / Full publication date: Autumn, 1983 / Copyright © 1983 American Association for Public Opinion Research.

Hobcraft, J. and Sacker, A. (2012). Guest editorial: the origins of understanding society. *Longitudinal and Life Course Studies*, 3(1):1–4.

Hochstim, J. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association*, 62(319):976–989.

Holbrook, A., Green, M., and Krosnick, J. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, 67(1):79–125.

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.

Holtgraves, T. (2004). Social desirability and Self-Reports: testing models of socially desirable responding. *Personality and Social Psychology Bulletin*, 30(2):161–172.

Hox, J. J., De Leeuw, E. D., and Zijlmans, E. A. O. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology*, 6.

Hu, L.-t. and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55.

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.

Jäckle, A., Roberts, C., and Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, 78(1):3–20.

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 nonexperimental panel studies? consistency of an attitude toward guestworkers. *Sociological Methods & Research*, 15(3):259–302.

Kalton, G. and Citro, C. (1995). Panel surveys: Adding the fourth dimension. *Innovation: The European Journal of Social Science Research*, 8(1):25–39.

Kankaraš, M. and Moors, G. (2010). Researching measurement equivalence in Cross-Cultural studies. *Psihologija*, 43(2):121–136.

Kankaraš, M., Vermunt, J., and Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods and Research*, (20):1–31.

Kaplowitz, M. D., Hadlock, T. D., and Levine, R. (2004). A comparison of web and mail survey response rates. *Public Opinion Quarterly*, 68(1):94–101.

Kaplowitz, M. D., Lupi, F., Couper, M. P., and Thorp, L. (2012). The effect of invitation design on web survey response rates. *Social Science Computer Review*, 30(3):339–349.

Kenny, D. A., Kaniskan, B., and McCoach, D. B. (2014). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, page 0049124114543236.

Kessler, R. C. and Greenberg, D. F. (1981). *Linear panel analysis: models of quantitative change*. Academic Press.

Klausch, T., Hox, J., and Schouten, B. (2015). Selection error in single-and mixed mode surveys of the dutch general population. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Klausch, T., Hox, J. J., and Schouten, B. (2013). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research*, 42(3):227–263.

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

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.

Link, M. and Mokdad, A. (2006). Can web and mail survey modes improve participation in a RDD-Based national health surveillance? *Journal of Official Statistics*, 22(2):293–312.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley Publishing Company, Inc.

Lugtig, P., Das, M., and Scherpenzeel, A. C. (2014). Nonresponse and attrition in a probability-based online panel for the general population. In Callegaro, M., editor, *Online panel research: a data quality perspective*, pages 135–153. Wiley.

Lugtig, P. J., Lensvelt-Mulders, G. J., 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. (2008). The problem of nonresponse. In Leeuw, E. D. d., Hox, J. J., and Dillman, D., editors, *International Handbook of Survey Methodology*, pages 35–55. Routledge Academic, 1 edition.

Lynn, P. (2009). Sample design for understanding society. *Understanding Society Working Paper Series*, (2009-01):1–46.

Lynn, P. (2013). Alternative sequential Mixed-Mode designs: Effects on attrition rates, attrition bias, and costs. *Journal of Survey Statistics and Methodology*, 1(2):183–205.

Lynn, P. (2014a). Targeted initial letters to longitudinal survey sample members: Effects on response rate, response speed, and sample composition. Understanding Society Working Paper 2014-08, University of Essex, Colchester.

Lynn, P. (2014b). Targeted response inducement strategies on longitudinal surveys. In Engel, U., editor, *Improving Survey Methods: Lessons from Recent Research*, pages 322–338. Routledge, New York.

Lynn, P., Hope, S., Jäckle, A., Campanelli, P., and Nicolaas, G. (2012). Effects of visual and aural communication of categorical response options on answers to survey questions. *ISER Working Paper Series*, (2012-21):1–31.

Lynn, P. and Kaminska, O. (2013). The impact of mobile phones on survey measurement error. *Public Opinion Quarterly*, 77(2):586–605.

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., and Vehovar, V. (2008). Web surveys versus other survey modes: a meta-analysis comparing response rates. *International Journal of Market Research*, 50(1):79–104.

Martin, P. and Lynn, P. (2011a). The effects of mixed mode survey designs on simple and complex analyses. *ISER Working Paper*.

Martin, P. and Lynn, P. (2011b). The effects of mixed mode survey designs on simple and complex analyses. *ISER Working Paper*.

Maurischat, C., Herschbach, P., Peters, A., and Bullinger, M. (2008). Factorial validity of the short form 12 (SF-12) in patients with diabetes mellitus. *Psychology Science*, 50(1):7.

Mavletova, A. and Couper, M. P. (2013). Sensitive topics in PC web and mobile web surveys: Is there a difference? *Survey Research Methods*, 7(3):191–205.

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. (2013). Understanding society – the UK household longitudinal study, innovation panel, waves 1-5, user manual. *University of Essex, Colchester*, pages 1–66.

Mehta, R. and Sivadas, E. (1995). Comparing response rates and response content in mail versus electronic mail surveys. *Journal of the Market Research Society*, 37(4):429–439.

Merad, S. (2012). Introducing web collection in the UK LFS. In *Data Collection for Social Surveys using Multiple Modes*, Wiesbaden.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, 29(2):177–185.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4):525–543.

Millar, M. M. and Dillman, D. A. (2011). Improving response to web and Mixed-Mode surveys. *Public Opinion Quarterly*, 75(2):249–269.

Millsap, R. E. (2012). *Statistical Approaches to Measurement Invariance*. Routledge Academic, 1 edition edition.

Millsap, R. E. and Yun-Tein, J. (2004). Assessing factorial invariance in Ordered-Categorical measures. *Multivariate Behavioral Research*, 39(3):479–515.

Morgan, S. L. and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, New York, 1 edition edition.

Moum, T. (1998). Mode of administration and interviewer effects in self-reported symptoms of anxiety and depression. *Social Indicators Research*, 45(1-3):279–318.

Muñoz-Leiva, F., Sánchez-Fernández, J., Montoro-Ríos, F., and Ibáñez-Zapata, J. A. (2009). Improving the response rate and quality in web-based surveys through the personalization and frequency of reminder mailings. *Quality & Quantity*, 44(5):1037–1052.

Muthén, B. and Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-Group and growth modeling in mplus. *Mplus Web Notes*, (4):1–22.

Muthén, B., du Toit, S., and Spisic, D. (1997). Robust inference using wighted least squares and quadratic estimation equations in latent variable modeling with categorical and continuous outcomes. *Technical Report*, pages 1–49.

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

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

Olson, K., Smyth, J. D., and Wood, H. M. (2012). Does giving people their preferred survey mode actually increase survey participation rates? an experimental examination. *Public Opinion Quarterly*, 76(4):611–635.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, U.K. ; New York, 2Rev e. edition edition.

Plewis, I. (1985). *Analysing change: measurement and explanation using longitudinal data*. J. Wiley.

Porter, S. R. and Whitcomb, M. E. (2007). Mixed-Mode contacts in web surveys: Paper is not necessarily better. *Public Opinion Quarterly*, 71(4):635–648.

Presser, S. and Stinson, L. (1998). Data collection mode and social desirability bias in self-reported religious attendance. *American Sociological Review*, page 137–145.

Resnick, B. and Nahm, E. (2001). Reliability and validity testing of the revised 12-item Short-Form health survey in older adults. *Journal of Nursing Measurement*, 9(2):151–161.

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

Révilla, M. (2012). Impact of the mode of data collection on the quality of survey questions depending on respondents' characteristics. *Bulletin of Sociological Methodology*, 116(1):44–60.

Révilla, M. and Saris, W. E. (2010). A comparison of the quality of ESS questions in different data collection modes. *RECSM Working Paper*, pages 1–31.

Révilla, M. A. (2013). Measurement invariance and quality of composite scores in a face-to-face and a web survey. *Survey Research Methods*, 7(1):17–28.

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

Rohani, C., Abedi, H. A., and Langius, A. (2010). The iranian SF-12 health survey version 2 (SF-12v2): factorial and convergent validity, internal consistency and test-retest in a healthy sample. *Iranian Rehabilitation Journal*, 8(12):4–14.

Salyers, M. P., Bosworth, H. B., Swanson, J. W., Lamb-Pagone, J., and Osher, F. C. (2000). Reliability and validity of the SF-12 health survey among people with severe mental illness. *Medical Care*, 38(11):1141–1150.

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

Saris, W. and Gallhofer, I. (2007b). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey research methods*, 1(1):29–43.

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

Saris, W. E., Satorra, A., and Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, 16(4):561–582.

Satorra, A. and Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4):507–514.

Schaefer, D. R. and Dillman, D. A. (1998). Development of a standard e-mail methodology: Results of an experiment. *Public opinion quarterly*, page 378–397.

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., and Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*, 42(6):1555–1570.

Schwarz, N., Hippler, H., and Noelle-Neumann, E. (1992). A cognitive model of response order effects in survey measurement. In Schwarz, N. and Sudman, S., editors, *Context effects in social and psychological research*, pages 187–201. Springer-Verlag, New York.

Schwarz, N., Strack, F., Hippler, H. J., and Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology*, 5(3):193–212.

Sharp, L. M. and Frankel, J. (1983). Respondent burden: A test of some common assumptions. *Public Opinion Quarterly*, 47(1):36–53.

Steenkamp, J. E. M. and Baumgartner, H. (1998). Assessing measurement invariance in Cross-National consumer research. *Journal of Consumer Research*, 25(1):78–107.

Steffick, D. (2000). Documentation of affective functioning measures in the health and retirement study. Technical report, Health and Retirement Study, Ann Arbor,MI.

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.

Sudman, S., Bradburn, N. M., and Schwarz, N. (1996). *Thinking about answers: the application of cognitive processes to survey methodology*. Jossey-Bass Publishers, San Francisco.

Sykes, W. and Collins, M. (1988). Effects of mode of interview: Experiments in the UK. In Groves, R., Biemer, P., Lyberg, L., Massey, J., Nicholls II, W., and Waksberg, J., editors, *Telephone Survey Methodology*, Wiley Series in Probability and Mathematical Statistics, pages 301–320. John Wiley & Sons, New York.

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

Uhrig, N. (2011). Using experiments to guide decision making in understanding society: Introducing the innovation panel. In McFall, S. and Garrington, C., editors, *Understanding Society: Early Findings from the First Wave of the UK's Household Longitudinal Study*. Colchester: University of Essex.

Uhrig, S. N. (2008). The nature and causes of attrition in the british household panel survey. *ISER Working Paper*, (2008-05):1–85.

van de Pol, F. and Langeheine, R. (1990). Mixed markov latent class models. In Clogg, C. C., editor, *Sociological methodology*, volume 20, pages 213–247. Blackwell, Oxford.

van de Schoot, R., Lugtig, P., and Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9(4):486–492.

Van de Vijver, F. (2003). Bias and equivalence: Cross-Cultural perspectives. In Harkness, J., Van de Vijver, F., and Mohler, P., editors, *Cross-cultural survey methods*, pages 143–155. J. Wiley, Hoboken, N.J.

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., Loosveldt, G., and Molenberghs, G. (2014). Evaluating mode effects in Mixed-Mode survey data using covariate adjustment models. *Journal of Official Statistics*, 30(1):1–21.

Vannieuwenhuyze, J. T. A., Loosveldt, G., and 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. T. A. and Révilla, M. (2013). Relative mode effects on data quality in Mixed-Mode surveys by an instrumental variable. *Survey Research Methods*, 7(3):157–168.

Visser, P., Krosnick, J., Marquette, J., and Curtin, M. (2000). Improving election forecasting: Allocation of undecided respondents, identification of likely voters, and response order effects. In Lavrakas, P. and Traugott, M., editors, *Election Polls, the News Media, and Democracy*. Chatham House, New York, 1St edition edition edition.

Voogt, R. and Saris, W. (2005). Mixed mode designs: Finding the balance between nonresponse bias and mode effects. *Journal of official Statistics*, 21(3):367–387.

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

Watson, N. and Wooden, M. (2009). Identifying factors affecting longitudinal survey response. In Lynn, P., editor, *Methodology of longitudinal surveys*, pages 157–181. Wiley, Chichester.

Weeks, M. F., Kulka, R. A., Lessler, J. T., and Whitmore, R. W. (1983). Personal versus telephone surveys for collecting household health data at the local level. *American Journal of Public Health*, 73(12):1389–1394.

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

Wood, M. and Kunz, S. (2014). CAWI in a mixed mode longitudinal design. Technical report, Understanding Society at the Institute for Social and Economic Research.

Appendix A

Item wording

SF1. In general, would you say your health is?

Excellent

Very good

Good

Fair

Poor

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?

SF2a. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf.

SF2b. Climbing several flights of stairs.

Yes, limited a lot

Yes, limited a little

No, not limited at all

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?

SF3a. Accomplished less than you would like.

SF3b. Were limited in the kind of work or other activities.

All of the time

Most of the time

Some of the time

A little of the time

None of the time

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)?

SF4a. Accomplished less than you would like.

SF4b. Did work or other activities less carefully than usual.

All of the time

Most of the time

Some of the time

A little of the time

None of the time

SF5. During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?

Not at all

A little bit

Moderately

Quite a bit

Extremely

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...

SF6a. Have you felt calm and peaceful?

SF6b. Did you have a lot of energy?

SF6c. Have you felt downhearted and depressed?

All of the time

Most of the time

Some of the time

A little of the time

None of the time

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 with friends, relatives, etc.)?

All of the time

Most of the time

Some of the time

A little of the time

None of the time

Appendix B

Tables

Table B.1: Estimates of individual change are equal across the two mode designs for eight out of twelve SF12 items.

Variable	Model	χ^2	df	RMSEA	CFI	$\Delta\chi^2$	df	p
SF1	Baseline by groups	104.66	30	0.053	0.995			
	Equal mean of slope	81.63	31	0.043	0.996	0.01	1	0.92
	Equal variance of slope	81.893	32	0.042	0.997	1	1	0.32
	Equal correlation	78.216	33	0.039	0.997	2.78	1	0.10
SF2a	Baseline by groups	55.095	16	0.052	0.994			
	Equal mean of slope	54.274	17	0.05	0.994	0.03	1	0.25
	Equal variance of slope	51.408	18	0.046	0.995	1	1	0.64
	Equal correlation	41.072	19	0.036	0.997	1.05	1	0.90
SF2b	Baseline by groups	47.637	16	0.047	0.996			
	Equal mean of slope	46.567	17	0.044	0.997	0.17	1	0.68
	Equal variance of slope	44.856	18	0.041	0.997	0.19	1	0.66
	Equal correlation	36.992	19	0.033	0.998	1.12	1	0.29
SF3a	Baseline by groups	91.3	30	0.048	0.983			
	Equal mean of slope	86.036	31	0.045	0.985	1.34	1	0.25
	Equal variance of slope	85.085	32	0.043	0.985	0.22	1	0.64
	Equal correlation	68.571	33	0.035	0.99	0.02	1	0.90
SF3b	Baseline by groups	84.511	30	0.045	0.988			
	Equal mean of slope	81.492	31	0.043	0.989	1.74	1	0.19
	Equal variance of slope	80.63	32	0.041	0.99	1.32	1	0.25
	Equal correlation	62.981	33	0.032	0.994	1.06	1	0.30
SF4a	Baseline by groups	95.329	30	0.049	0.958			
	Equal mean of slope	92.135	31	0.047	0.961	0.08	1	0.78
	Equal variance of slope	92.148	32	0.046	0.962	1.19	1	0.28
	Equal correlation	77.391	33	0.039	0.972	1.1	1	0.30
SF4b	Baseline by groups	68.638	30	0.038	0.962			
	Equal mean of slope	68.901	31	0.037	0.963	2.19	1	0.14
	Equal variance of slope	68.28	32	0.036	0.965	0.45	1	0.50
	Equal correlation	60.74	33	0.031	0.973	1.11	1	0.29
SF5	Baseline by groups	65.812	30	0.037	0.987			
	Equal mean of slope	62.807	31	0.034	0.988	0.47	1	0.49
	Equal variance of slope	62.107	32	0.032	0.989	1.1	1	0.29
	Equal correlation	52.172	33	0.025	0.993	0.08	1	0.78

Table B.2: Descriptive statistics from balanced sample of Health and Retirement Study

		Telephone	Face to face	Web	Total sample
CESD					
Depressed	No	92.37	92.96	89.70	91.67
	Yes	6.40	5.81	9.78	7.33
	Missing	1.23	1.23	0.52	0.99
Everything an effort	No	86.83	87.08	85.14	86.35
	Yes	11.90	11.69	14.00	12.53
	Missing	1.26	1.23	0.52	1.12
Restless sleep	No	72.96	73.58	72.22	72.92
	Yes	25.81	25.19	27.25	26.08
	Missing	1.23	1.23	0.52	0.99
Happy	No	10.43	9.57	13.81	11.27
	Yes	88.28	89.11	85.54	87.64
	Missing	1.29	1.32	0.65	1.09
Lonely	No	88.71	89.36	87.57	88.55
	Yes	10.03	9.38	11.93	10.45
	Missing	1.26	1.26	0.49	1.00
Enjoyed life	No	6.06	5.38	11.47	7.64
	Yes	92.68	93.29	87.82	91.26
	Missing	1.26	1.32	0.71	1.10
Felt sad	No	86.19	86.16	85.73	86.02
	Yes	12.52	12.49	13.60	12.87
	Missing	1.29	1.35	0.68	1.11
Could not get going	No	85.88	85.88	83.17	84.98
	Yes	12.86	12.86	16.15	13.95
	Missing	1.26	1.26	0.68	1.07
Had a lot of energy	No	39.28	40.23	43.34	40.95
	Yes	59.37	58.26	55.58	57.74
	Missing	1.35	1.51	1.08	1.31

Table B.2: Descriptive statistics from balanced sample of Health and Retirement Study

		Telephone	Face to face	Web	Total sample
Mild activity	More than once a week	66.87	65.83	68.04	66.91
	Once a week	22.45	23.19	18.58	21.41
	One to three times a month	6.43	6.18	8.46	7.02
	Hardly ever or never	4.24	4.74	4.15	4.38
	Missing	0.00	0.06	0.77	0.28
Moderate activity	More than once a week	58.54	58.35	57.46	58.12
	Once a week	16.24	16.79	14.24	15.76
	One to three times a month	11.04	10.74	14.83	12.20
	Hardly ever or never	14.12	14.12	13.07	13.77
	Missing	0.06	0.00	0.40	0.15
Vigorous activity	More than once a week	32.17	31.50	33.84	33.84
	Once a week	11.53	11.75	11.10	11.46
	One to three times a month	11.17	11.44	17.26	13.29
	Hardly ever or never	44.97	45.06	37.50	45.51
	Missing	0.15	0.25	0.31	0.24
Importance of religion	Very important	58.17	58.75	55.98	57.63
	Somewhat important	23.01	22.67	23.01	22.90
	Not too important	18.67	18.49	20.64	19.27
	Missing	0.15	0.09	0.37	0.21
How often do you go to religious service?	More than once a week	13.78	14.12	15.32	14.41
	Once a week	24.33	23.69	22.96	23.66
	Two or three times a week	11.38	11.26	9.57	10.74
	One or more times a year	23.47	22.81	22.59	23.62
	Not at all	26.79	27.10	28.30	27.40
	Missing	0.25	0.03	0.25	0.17

Table B.3: Descriptive statistics Innovation Panel 5

		Freq.	Percent
Full response	No	742	29.42
	Yes	1780	70.58
Mode treatment	F2F	857	33.98
	Web/F2F	1665	66.02
Email	No	1176	46.63
	Yes	1346	53.37
Partner's email	No	1684	66.77
	Yes	838	33.23
Education	Degree	576	22.84
	A levels	237	9.4
	GCSE or CSE	752	29.82
	Vocational/none	941	37.31
	Missing	16	0.63
Urban	No	612	24.27
	Yes	1910	75.73
Female	No	1158	45.92
	Yes	1364	54.08
Partner	No	946	37.51
	Yes	1576	62.49
White British	No	359	14.23
	Yes	2163	85.77
Employed	No	1118	44.33
	Yes	1404	55.67
Own house	No	657	26.05
	Yes	1865	73.95
Has mobile	No	195	7.73
	Yes	2327	92.27
Has broadband	No	502	19.9
	Yes	2020	80.1
Uses internet daily	No	1449	57.45
	Yes	1073	42.55
Would not answer by web	No	1861	73.79
	Yes	661	26.21
	Mean	**Std. Dev**	**Max**
Hh. Size	2.84	1.44	10
Age	48.185	18.15	65