Understanding Society

Working Paper Series

No. 2014 – 05

July 2014

Impact of mixed modes on measurement errors and estimates of change in panel data

Alexandru Cernat

Institute for Social and Economic Research

University of Essex

Understanding Society
THE UK HOUSEHOLD LONGITUDINAL STUDY

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

Non-Technical Summary

Survey questionnaires can be administered through various mediums, from face-to-face interviews to self-administration on the internet. This opens up the possibility of mixing two or more mediums both for the same respondent (e.g., at different points in time) and across individuals (e.g., offering the possibility of answering by web to some respondents). These decisions influence the quality of the data collected.

In this study I compare a design that applies questionnaires face-to-face to one that uses a combination of telephone and face-to-face in the Understanding Society Innovation Panel. By comparing the results of the SF12, a health scale, I will show the impact of changing the mode design on data quality and estimates of individual change.

Results show that the two designs, single mode face-to-face and multi mode telephone/face-to-face, lead to similar data quality estimates for the 12 items analyzed. Differences appear for one variable in two out the four waves. On the other hand, the degree of individual change is overestimated for three out of 12 items when the mixed mode (telephone/face-to-face) design was used. I conclude that combining telephone and face-to-face may lead to similar data quality to a single mode face-to-face but the change to such a design may overestimate individual change in panel data.

Impact of mixed modes on measurement errors and estimates of change in panel data

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

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

Abstract

Mixed modes (MM) 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 MM on panel data I 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 MM. Also, three of the SF12 items overestimate variance of change in time in the MM design. I conclude that using a MM design leads to minor measurement differences but it can result in the overestimation of individual change compared to a single mode approach.

Key words: longitudinal survey, mixed mode survey, CAPI, CATI, data quality, latent measurement models, equivalence.

JEL Codes: C81, C83

Acknowledgements: I would like to thank all the people that helped with this paper: Peter Lynn, Nick Allum, Peter Lugtig, Jorre Vannieuwenhuyze and Oana Mihai. This work was supported by a +3 PhD grant awarded by the UK Economic and Social Research Council.

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

As more surveys are moving to or taking into consideration MMDs additional research regarding their effects on selection, measurement and statistical estimates is needed in order to make informed decisions. 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 MMDs 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 I 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 MMD is implemented and in subsequent waves in order to estimate both the direct and the lasting effects due to mode design. Secondly, I 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 MMDs 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 I will use the latter method by taking advantage of the randomization into mode design in the UKHLS-IP. Thus, the results will compare mixed modes (sequential CATI–CAPI) to a CAPI single mode design (SMD), 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 their mixture. 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.

2 Background

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

2.1 Mode differences: CAPI vs. CATI

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 De Leeuw and van der Zouwen, 1988; Groves and Kahn, 1979; Groves, 1990; Schwarz et al., 1991, for an overview). Due to the data collection design used here I 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 subsection 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 to 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., Bishop and Smith, 2001; Holbrook et al., 2007; McClendon, 1991) 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).

Another aspect that differentiates the 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 differences between CATI and CAPI are not as large, the latter being advantaged by 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 (Aquilino, 1992, 1998; Greenfield et al., 2000; Groves and Kahn, 1979; Hochstim, 1967; Holbrook et al., 2003; Jäckle et al., 2010)

Finally, satisficing (Krosnick, 1991), the tendency not to think thoroughly 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 and Kahn, 1979; Holbrook et al., 2003; Schwarz et al., 1991), 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 tend to have slightly bigger response rates and smaller response bias when compared to telephone surveys (Aquilino, 1992; Biemer, 2001; De Leeuw and van der Zouwen, 1988; Groves and Kahn, 1979; Voogt and Saris, 2005; Weeks et al., 1983). When analyzing effects on measurement most studies find small or no differences at all (Aquilino, 1998; De Leeuw and van der Zouwen, 1988; Greenfield et al., 2000), 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 MMDs 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 have found inconclusive results. For example, in one of the first mixed mode survey experiments Hochstim (1967) found no differences in response rates between two mixed mode designs that included telephone compared to a face-to-face, single mode, approach. On the other hand, other surveys found a better response rate when using a MMD (e.g., Voogt and Saris, 2005). Similarly, for measurement, Révilla (2010) shows that for some scales, such as social trust, there is no measurement difference between single and mixed modes 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 MMDs, concurrent and sequential (De Leeuw, 2005). Although these differences are found, they are nevertheless not as large as expected, being smaller than the differences between the methods used (Révilla, 2010).

2.2 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 MMD. 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 relationship 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 (Lugtig et al., 2014; Watson and Wooden, 2009). This phenomenon can be complicated when combined with a MMD. For example, Lynn (2013) has showed how two different MMDs that used a CATI-CAPI sequential approach can lead to different attrition patterns, both compared to each-other and to a CAPI SMD.

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 a gain in reliability and consistency in time of responses (Sturgis et al., 2009). Applying MMDs 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 MMD. Cernat (2013) 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 MMD. 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 research has showed that these change coefficients are less reliable than the variables that compose them (Kessler and Greenberg, 1981; Plewis, 1985). Their estimation is even more complicated in the case of longitudinal studies that either use a MMD from the beginning or change to such a design in time. Any differences confounded with the new mode(s) or the MMD will bias estimates of change in unknown ways. So far there is no research on this topic.

3 Data and methodology

In order to tackle some of the issues presented above I will be analyzing the first four waves of UKHLS-IP, which is described in more detail in the next subsection. The statistical approach used, equivalence testing of the measurement model and the estimates of change, as well as the procedures applied are discussed in the subsequent sections.

3.1 Data

In order to investigate the impact of mixing modes on errors and estimates of change in panel data I 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 rates for these were 54.8% (465 households) while the individual ones were 80.1% (for more details: McFall et al., 2013) .

The UKHLS-IP was developed in order 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 SMD in the second wave. Approximately 68% of the respondents in the MMD responded by telephone, while the rest did so face-to-face (McFall et al., 2013). Overall, the response rates for the MMD were significantly lower than

Table 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 the SMD (see for mode details: Lynn, 2013). For the rest of the four waves all respondents participated using CAPI SMD.

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 MMD on measurement errors and estimates of change the SF12 will be analyzed. This scale is the short version of the SF32 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 by each item can be found in Table 1.

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).

3.2 Equivalence testing

The previous section has revealed that the main focus of the mixed modes research is to find causal effect of mode or mode design systems. This can be done either with specific statistical models or (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 I am able to compare the SMD to the MMD without having to use statistical models for selection. The remaining task is to compare the two mode designs. In order to do this I will be 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 in 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, 2007), social desirability (Holtgraves, 2004; Tourangeau et al., 2000) or acquiscence (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 bias (van de Vijver, 2003). This can be incorporated in a Multi Group Confirmatory Factor Analysis when comparing more groups using equivalence (Millsap, 2012; Steenkamp and Baumgartner, 1998; van de Schoot et al., 2012).

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 equal across the groups. Furthermore, the model can also assume that the intercepts are equal across groups, leading to scalar equivalence. This latter 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 substantial latent variable. The measurement model can also be conceptualized as one composed of three parts: random error, systematic error and the substantial 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 in the different parts of the measurement model. 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, I am able to investigate how much the switch in data collection approach biases individual estimates of change.

3.3 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 substantial 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, I do not expect any differences at the starting point between the two groups because the same mode design was applied, and selection in 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 I will be using Mplus 7 (Muthén and Muthén, 2012). For missing data Full Information estimation will be used, assuming MAR given the measurement model (Enders, 2010). For the equivalence testing Weighted Least Squares Means and Variance (WLSMV, Asparouhov and Muthén, 2010; Millsap and Yun-Tein, 2004; Muthén et al., 1997) will be applied in order to take into account the categorical character of the data. No weighting will be used.

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 I will use the Theta approach (Millsap and Yun-Tein, 2004; Muthén and Asparouhov, 2002). 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 see Millsap and Yun-Tein (2004), Millsap (2012) and Muthén and Asparouhov (2002).

4 Analysis and Results

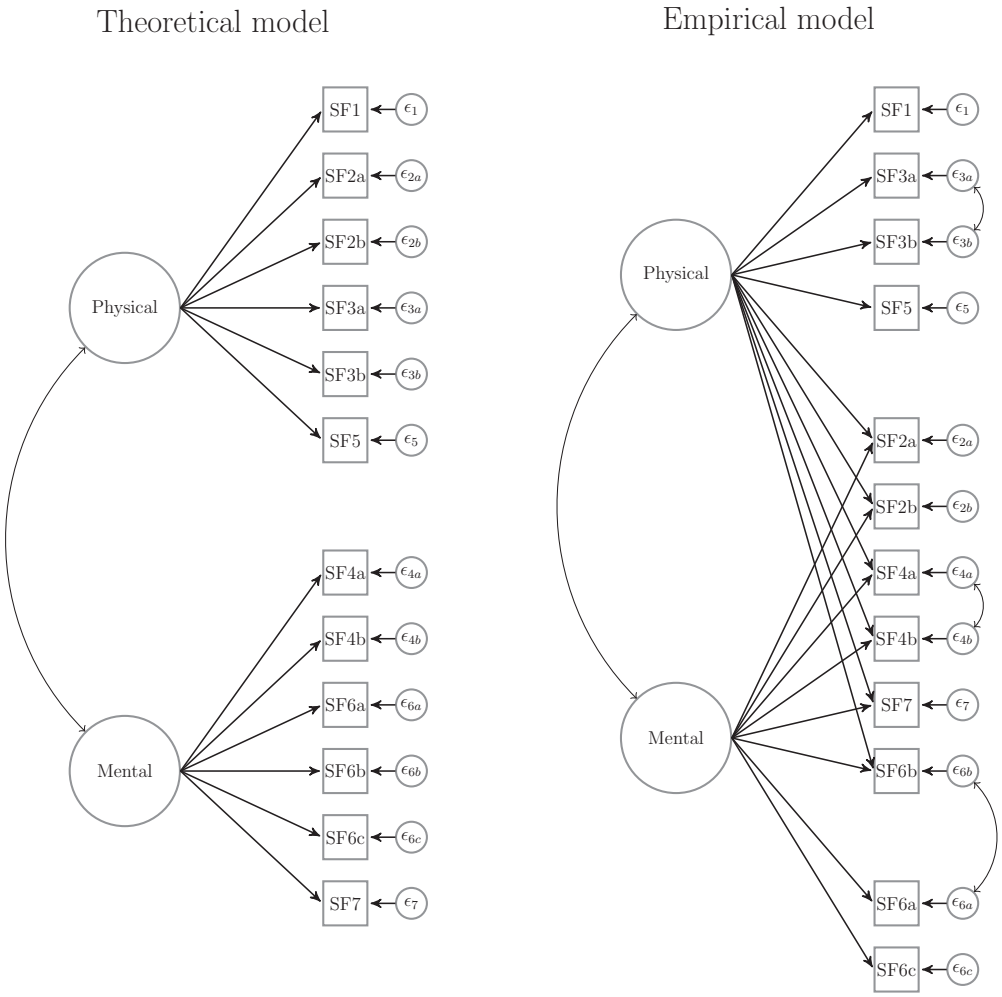
In order to evaluate the effect of the MMD on data quality and estimates of change, three steps will be followed. In the first step, presented in the next subsection, I will analyze the measurement model for the SF12 scale. This will be used in subsection 4.2 to identify how the MMD influence data quality both in the wave in which applied, namely wave two of the UKHLS-IP, and in the following waves. In the last subsection, I will show what is the effect of changing to a MMD on the estimates of individual change in health.

4.1 Base model

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 Figure 1. As opposed to the SF32, the subdimensions are only measured by one or two variables (see Table 1) and, thus, are not reliable enough to be estimated individually. As a result, the two

Figure 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.



main dimensions, physical and mental health, will be estimated using latent variables, each with six variables.

This is the first model tested and presented in Table 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, I add cross-loadings and correlated errors. 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[1]. All the new relationships lead to improvements in fit. The final model (which is also presented in Figure 1) has a good fit both for RMSEA (0.033) and CFI (0.998) and also fits well in waves two, three and four.

[1]The Mplus function DIFFTEST is used here and the next sections for the $\Delta\chi^2$ method, because of the WLMSV estimation. For more details refer to: http://www.statmodel.com/chidiff.shtml

Table 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.

Model	χ^2	df	RMSEA	CFI	$\Delta\chi^2$	Δdf	p
Ware et al. 2007	1493.63	53	0.103	0.977			
SF6b	1143.05	52	0.09	0.983	158.83	1	0.00
SF7	746.91	51	0.073	0.989	149.79	1	0.00
SF3b with SF3a	592.93	50	0.065	0.992	86.13	1	0.00
SF4b	474.69	49	0.058	0.993	59.88	1	0.00
SF4a with SF4b	418.94	48	0.055	0.994	38.86	1	0.00
SF4a	306.81	47	0.046	0.996	52.03	1	0.00
SF6b with SF6a	230.31	46	0.04	0.997	74.15	1	0.00
SF2b	199.14	45	0.037	0.998	22.73	1	0.00
SF2a	170.24	44	0.033	0.998	19.66	1	0.00
Wave 2	134.75	44	0.033	0.998			
Wave 3	159.45	44	0.043	0.998			
Wave 4	214.18	44	0.045	0.997			

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 (Cernin et al., 2010; Resnick and Nahm, 2001; Rohani et al., 2010; Salyers et al., 2000). 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.

4.2 Equivalence testing across the four waves

Using the model chosen in the previous subsection (empirical model in Figure 1) I 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 shows the results of this analysis. Using once again the $\Delta\chi^2$ method, it can be seen that all constraints hold in wave one of the data, meaning that the measurement model is compeltely 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. On the other hand, scalar equivalence, equal

thresholds, is not reached as the $\Delta\chi^2$ is significant. 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, so, 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 the threshold for one of the variables is not equal across the two mode designs. This time the SF4b, 'Did work less carefully', is unequal. 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 4). As mentioned before, this difference 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 MMD, with the use of CATI, leads to lower social desirability. 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 4). 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., Sturgis et al., 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 in these data.

4.3 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. The results indicate that only three variables have any differences in their estimations of individual change (Table 5): SF6a ('Felt calm and peaceful'), SF6c ('Felt downhearted and depressed') and SF6b ('Lot of energy'). The first two are part of the same subdimension, mental health, while SF6b measures the vitality subdimension. All three 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).

Table 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
	Baseline by groups	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
Wave 1	**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
	Baseline by groups	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
Wave 2	**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
	Baseline by groups	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
Wave 3	**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
	Baseline by groups	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
Wave 4	**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 freely estimated coefficients.

Table 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

A more detailed look indicates that the mixed mode design leads to the overestimation of individual change for all three variables: 0.116 versus 0.047 for SF6a, 0.078 versus 0.025 for SF6b and 0.108 versus 0.017 for SF6c. 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 4.2 and previous research on this data (Cernat, 2013). Lastly, attrition may cause a mode design effect that also impacts estimates of change. Previous research (Lynn, 2013) appears to support such a hypothesis.

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 MMD (in case of acquiescence), or the SMD (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 takes place 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 only three of the SF12 variables have mode design effects in their estimates of individual change. For all three of them the same coefficient is biased in the same direction. It appears that for these items the mixed mode

Table 5: For three 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 groups**	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 groups**	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 groups**	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

Gray background indicates unequal coefficients.

design overestimates variation of individual change. All three 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 lasting. On the other hand, the mode design effects seem to decrease/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 only in three 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 SEM. Both of them can be used in longitudinal studies and, combined with a quasi-experimental design, they make for strong tools to be used in this field. Equivalence testing with CFA has two more advantages. Firstly, it can also be used in

cross-sectional designs, such as those used by the European Social Survey mode experiments (Martin, 2011). Secondly, and more importantly, it can be used to correct for mode or mode design effects when two or more such designs are used. Thus, using the analysis presented above it would be possible to save the scores of the mental and health dimensions while taking into account the measurement differences in mode designs. These can be added to the dataset and made available to users. As more surveys are looking to switch to MMDs, this may prove a useful way to correct for measurement differences when scales are used.

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 SMD 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 SMD after just one wave of using the MMD.

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 and Loosveldt, 2012). Here only theoretical arguments and previous empirical work are explored as potential explanations.

References

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

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.

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

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.

Cernat, A. (2013). The impact of mixing modes on reliability in longitudinal studies. *ISER Working Paper*, (09):1–27.

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.

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.

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

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

De Leeuw, E. D. and 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.

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

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. 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.

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.

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. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, 78(1):3–20.

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

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.

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

Martin, P. (2011). What makes a good mix? chances and challenges of mixed mode data collection in the ESS. *London: Centre for Comparative Social Surveys, City University*, (Working Paper No. 02).

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.

McClendon, M. J. (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.

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

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.

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

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

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.

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

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.

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

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.

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

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. A., van de Vijver, F., and Mohler, P., editors, *Cross-cultural survey methods*, pages 143–155. J. Wiley, Hoboken, N.J.

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

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*. 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.

6 Annex

Table 6: Estimates of individual change are equal across the two mode designs in nine 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
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