Survey Research Methods (2015)
Vol. 9, No. 2, pp. 83-99
doi:10.18148/srm/2015.v9i2.5851

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

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

Alexandru Cernat
Institute for Social and Economic Research
University of Essex
United Kingdom

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.

Keywords: mixed modes; measurement model; latent growth models; panel data

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. Although these are exciting opportunities for innovation in survey methodology they also provide a number of unique challenges.

Some of these challenges 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, Burton, Jäckle, Lynn, & Uhrig, 2013). The Short Form 12-item Survey (SF12) health scale (Ware, Kosinski, Turner-Bowker, & Gandek, 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 randomization 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 ef-*

Alexandru Cernat, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ UK (cernat@essex.ac.uk)

fects, 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.

2 Background

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 & van der Zouwen, 1988; R. M. Groves, 1990; R. Groves & Kahn, 1979; Schwarz, Strack, Hippler, & Bishop, 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.

An important aspect that differentiates the two modes is the perceived legitimacy of the survey (Tourangeau, Rips, & Rasinski, 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 (R. M. 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 (W. S. Aquilino, 1992; W. Aquilino, 1998; Greenfield, Midanik, & Rogers, 2000; R. Groves & Kahn, 1979; Hochstim, 1967; Holbrook, Green, & Krosnick, 2003; Jäckle, Roberts, & Lynn, 2010)

Additionally, satisficing (J. 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 (R. Groves & 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 (J. Krosnick, 1991) and has been found in previous research focused on comparing the two modes (e.g., Holbrook et al., 2003; J. A. Krosnick, Narayan, & Smith, 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 (W. S. Aquilino, 1992; Biemer, 2001; De Leeuw & van der Zouwen, 1988; R. Groves & Kahn, 1979; Voogt & Willem Saris, 2005; Weeks, Kulka, Lessler, & Whitmore, 1983). When analyzing effects on measurement most studies find small or no differences at all (W. Aquilino, 1998; De Leeuw & 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 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 Willem 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 (2011) 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, Smyth, and Wood (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, Hox, and Schouten (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 M. Révilla (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 (M. Révilla, 2010). Similarly, Klausch, Hox, and Schouten (2013) have found signif-

icant differences in data quality between self-administered and interviewer modes but not between CAPI and CATI within a mixed mode survey.

2.1 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 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 (P. Lugtig, Das, & Scherpenzeel, 2014; Watson & Wooden, 2009). 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 (Sturgis, Allum, & Brunton-Smith, 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 (2014) 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 (Kessler & Greenberg, 1981; Plewis, 1985). 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 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 4 (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 in Table 1. For exact wording and response categories refer to the Appendix

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

Table 1
The SF12 scale measures physical and mental health and is based on eight initial subdimensions measured in SF32.

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

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 & McClendon, 2000).

3.1 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, K. 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 & Fiske, 1959; W. Saris, Satorra, & Coenders, 2004; W. E. Saris & Gallhofer, 2007), social desirability (Holtgraves, 2004; Tourangeau et al., 2000) or acquiescence (Billiet & Davidov, 2008; Billiet & 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 (Byrne, Shavelson, & Muthén, 1989; Meredith, 1993; Millsap, 2012; Steenkamp & Baumgartner, 1998; van de Schoot, Lugtig, & Hox, 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 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 (K. Bollen, 1989), the in-

tercept 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, K. A. Bollen & 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.

3.2 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 (L. Muthén & B.O. Muthén, 2012) with Weighted Least Squares Means and Variance (WLSMV, Asparouhov & Bengt Muthén, 2010; Millsap & Yun-Tein, 2004; Bengt Muthén, du Toit, & Spisic, 1997). 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 (Millsap & Yun-Tein, 2004; B.O. Muthén & Asparouhov, 2002). This implies adding the following constraints to the models in order to have convergence:

- all intercepts are fixed to 0;

[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.

- 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 B.O. Muthén and Asparouhov (2002).

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 Figure A1. As opposed to the SF36, 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 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 2.[2] The model has a moderate fit, with the CFI indicating good fit (Hu & 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 W. E. Saris, Satorra, and van der Veld (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 A1) 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 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, Herschbach, Peters, & Bullinger, 2008) or the proximity. Also, some of the cross-loadings found here were highlighted by previous research on the scale (Cernin, Cresci, Jankowski, & Lichtenberg, 2010; Resnick & Nahm, 2001; Rohani, Abedi, & Langius, 2010; Salyers, Bosworth, Swanson, Lamb-Pagone, & Osher, 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.

4.1 Equivalence testing across the four waves

Using the model chosen in the previous subsection (empirical model in Figure A1) we will test the cumulative constraints of the measurement model across the two mode designs using the sequence presented in Section 3.2. 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. 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

[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, Kaniskan, & McCoach, 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.

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.

								Misspecification				
Model	χ^2	df	RMSEA	CFI	$\Delta\chi^2$	Δdf	p	Coefficient	MI[a]	EPC[b]	Power	NCP[c]
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	−37.04	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								

[a] Modification Indice [b] Expected Parameter Change [c] Non-Centrality Parameter

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.

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 freely estimated coefficients.

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 (49.1 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 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 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 3.2), 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.

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.

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

In wave three, the thresholds of SF4b ('Did work less carefully') are significantly different between the two groups (Table 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., 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 which disappears by wave four.

4.2 Equivalence of latent growth models

Next, for each variable of the SF12, the LGM presented in Section 3.2 are tested using the $\Delta\chi^2$ method. For example, the Growth Model for SF6a (Table 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 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 A1). 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 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 4.1 and previous research on this data. For example, Cernat (2014) 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.

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, 2011; M. A. 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

Table 5
For three out of the 12 items tested the mixed mode design has significantly different variance of the slope.

Model	χ^2	df	RMSEA	CFI	$\Delta\chi^2$	df	p
Variable 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
Variable 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
Variable 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
Variable 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 unequal coefficients.

measurement and selection into mode (e.g., P. J. Lugtig, Lensvelt-Mulders, Frerichs, & Greven, 2011; Schouten, van den Brakel, Buelens, van der Laan, & Klausch, 2013; Vannieuwenhuyze & Loosveldt, 2012). 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.

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

Asparouhov, T. & Muthén, B. [Bengt]. (2010). Weighted least squares estimation with missing data. *Technical Report*, 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. & 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. doi:10.1177/0049124107313901

Billiet, J. & McClendon, M. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(4), 608–628. doi:10.1207/S15328007SEM0704_5

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

Bollen, K. A. & Curran, P. J. (2005). *Latent curve models: a structural equation perspective* (1st ed.). Wiley-Interscience.

Byrne, B., Shavelson, R., & 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. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod

matrix. *Psychological Bulletin*, *56*(2), 81–105. doi:10.1037/h0046016

Cernat, A. (2014). The impact of mixing modes on reliability in longitudinal studies. *Sociological Methods & Research*, (in press), 1–31.

Cernin, P. A., Cresci, K., Jankowski, T. B., & 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. doi:10.1891/1061-3749.18.1.49

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. doi:10.1037/a0013193

Couper, M. (2012). *Assesment of innovations in data collection technology for undersanding society*. Economic and Social Research Council. Retrieved from http://eprints.ncrm.ac.uk/2276/1/ESRC_Review_of_Mixed_Mode_Data_Collection_-_Household_Panel_Studies__(January_2012).pdf

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. & van der Zouwen, J. (1988). Data quality in telephone and face to face surveys: a comparative Meta-Analysis. In R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 283–299). Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons.

Greenfield, T. K., Midanik, L. T., & 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. M. (1990). Theories and methods of telephone surveys. *Annual Review of Sociology*, *16*(1), 221–240. doi:10.1146/annurev.so.16.080190.001253

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

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., & 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.

Holtgraves, T. (2004). Social desirability and Self-Reports: testing models of socially desirable responding. *Personality and Social Psychology Bulletin*, *30*(2), 161–172. doi:10.1177/0146167203259930

Hu, L.-t. & 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. doi:10.1080/10705519909540118

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, *78*(1), 3–20. doi:10.1111/j.1751-5823.2010.00102.x

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2014). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 0049124114543236. doi:10.1177/0049124114543236

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

Klausch, T., Hox, J. J., & Schouten, B. (2013). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research*, *42*(3), 227–263. doi:10.1177/0049124113500480

Klausch, T., Hox, J., & 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)*.

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

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

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

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

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

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. doi:10.1093/jssam/smt015

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).

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

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

McFall, S., Burton, J., Jäckle, A., Lynn, P., & Uhrig, N. (2013). Understanding society – the UK household longitudinal study, innovation panel, waves 1-5, user manual. *University of Essex, Colchester*, 1–66.

Merad, S. (2012). Introducing web collection in the UK LFS. In *Data collection for social surveys using multiple modes*. Wiesbaden.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. doi:10.1007/BF02294825

Millsap, R. E. (2012). *Statistical approaches to measurement invariance* (1 edition). Routledge Academic.

Millsap, R. E. & Yun-Tein, J. (2004). Assessing factorial invariance in Ordered-Categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515. doi:10.1207/S15327906MBR3903_4

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

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

Muthén, L. & Muthén, B. [B.O.]. (2012). *Mplus user's guide. seventh edition*. Los Angeles: CA: Muthén & Muthén.

Olson, K., Smyth, J. D., & 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. doi:10.1093/poq/nfs024

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

Resnick, B. & 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. 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.

Rohani, C., Abedi, H. A., & 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., & 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. [W.], Satorra, A., & Coenders, G. (2004). A new approach to evaluating the quality of measurement instruments: the Split-Ballot MTMM design. *Sociological Methodology*, *34*(1), 311–347.

Saris, W. E. & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research* (1st ed.). Wiley-Interscience.

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 561–582. doi:10.1080/10705510903203433

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., & Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*, *42*(6), 1555–1570. doi:10.1016/j.ssresearch.2013.07.005

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

Steenkamp, J. M. & Baumgartner, H. (1998). Assessing measurement invariance in Cross-National consumer research. *Journal of Consumer Research*, *25*(1), 78–107. doi:10.1086/209528

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

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response* (1st ed.). Cambridge University Press.

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. doi:10.1080/17405629.2012.686740

van de Vijver, F. (2003). Bias and equivalence: Cross-Cultural perspectives. In J. A. Harkness, F. van de Vijver, & P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 143–155). Hoboken, N.J.: J. Wiley.

Vannieuwenhuyze, J. T. A. & 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. doi:10.1177/0049124112464868

Voogt, R. & Saris, W. [Willem]. (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., & Gandek, B. (2007). *User's manual for the SF-12v2 health survey*. QualityMetric, Incorporated.

Watson, N. & Wooden, M. (2009). Identifying factors affecting longitudinal survey response. In P. Lynn (Ed.), *Methodology of longitudinal surveys*. Chichester: Wiley.

Weeks, M. F., Kulka, R. A., Lessler, J. T., & 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.

Appendix

Question 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

Table A1
Estimates of individual change are equal across the two mode designs for nine out of twelve SF12 items.

Model	χ^2	df	RMSEA	CFI	$\Delta\chi^2$	df	p
Variable 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
Variable 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
Variable 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
Variable 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
Variable 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
Variable 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
Variable 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
Variable 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

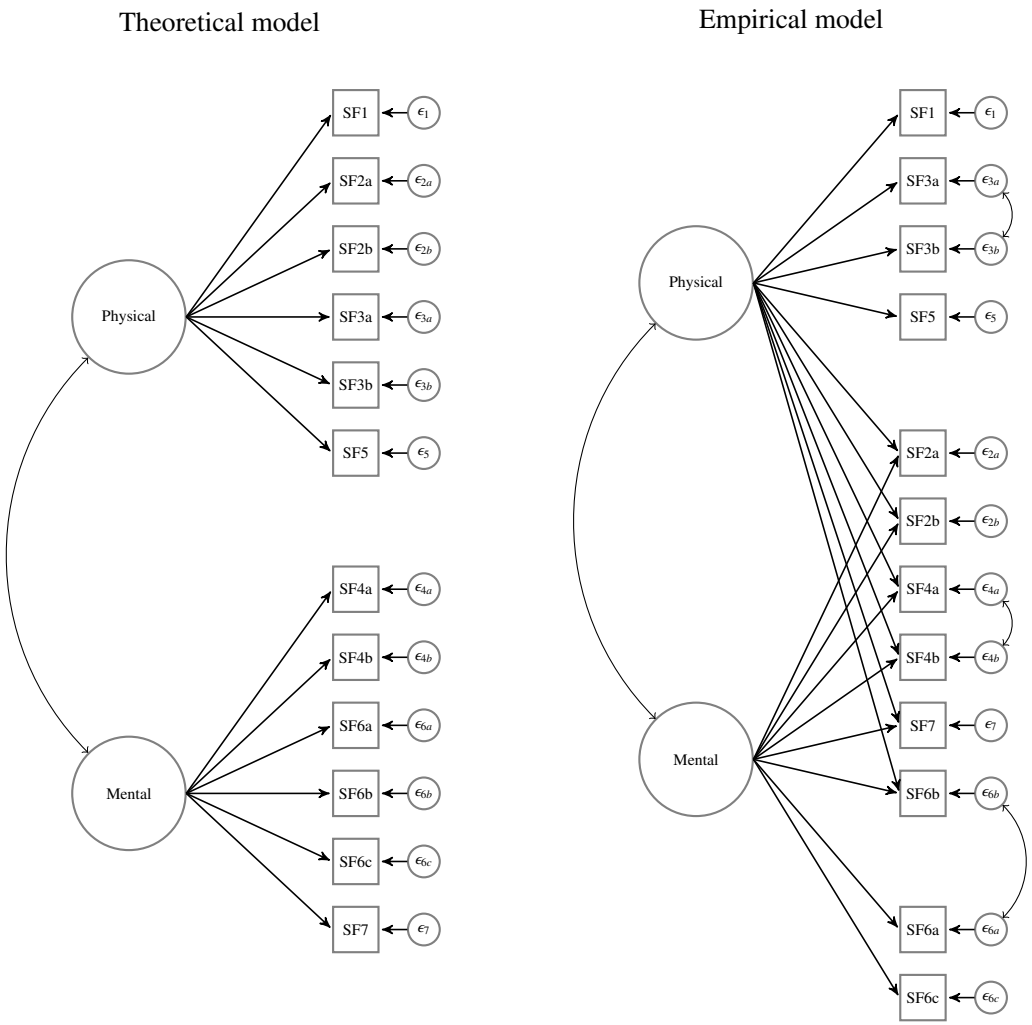

Figure A1. The theoretical model of the SF12 does not fit the UKHLS-IP data. A number of cross-loadings and correlated errors are evident in the data.