Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Surv Stat Methodol. Author manuscript; available in PMC 2017 June 11.

Published in final edited form as:

J Surv Stat Methodol. 2016 December ; 4(4): 501–524. doi:10.1093/jssam/smw021.

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

Alexandru Cernat,
The University of Manchester

Mick P. Couper, and
University of Michigan

Mary Beth Ofstedal
University of Michigan

Abstract

Using multiple modes to collect data is becoming a standard practice in survey agencies. While this should lower costs and reduce non-response error it may have detrimental effects on measurement quality. This is of special concern in panel surveys where a key focus is on measuring change over time and where changing modes may have an effect on key measures. In this paper we use a quasi-experimental design from the Health and Retirement Study to compare the measurement quality of two scales between face-to-face, telephone and Web modes. Panel members were randomly assigned to receive a telephone survey or enhanced face-to-face survey in the 2010 core wave, while this was reversed in the 2012 core wave. In 2011, panelists with Internet access completed a Web survey containing selected questions from the core waves. We examine the responses from 3251 respondents who participated in all three waves, using latent models to identify measurement mode effects. The two scales, depression and physical activity, show systematic differences between interviewer administered modes (i.e., face-to-face and telephone) and the self-administered one (i.e., Web). Possible explanations are discussed.[1]

1 Introduction

As surveys increasingly turn to mixed-mode designs, concerns about mode effects on measurement are being raised. And while mixed-mode strategies are often adopted for cost reasons, the trade-off in terms of measurement needs to be understood. This is especially true of panel studies where a key focus is on measuring change over time and a necessary assumption is measurement invariance over waves of data collection (Cernat, 2015b,a). Much of the research on mode effects has involved cross-sectional designs, with subjects randomly assigned to one mode of data collection or another. This often makes it hard to disentangle selection effects (those who choose to respond in a particular mode) from

[1]**Acknowledgements:** We thank Peter Lynn, Brady West, Oliver Lipps, and Hayk Gyuzalyan for their help with this paper; we also thank the two reviewers and the editor (Roger Tourangeau) for their helpful comments. This work was supported by a +3 PhD grant and an Overseas Institutional Visit grant awarded by the UK Economic and Social Research Council to the first author.

Author correspondence: alexandru.cernat@manchester.ac.uk.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

measurement effects. Changing modes in a panel study may similarly confound true change with effects of mode (Cernat, 2015a). The optimal experimental design for disentangling selection and measurement effects while controlling for temporal change would involve randomly assigning subjects to different modes at different times (e.g., in a randomized cross-over design). Such designs (e.g., Gmel, 2000; Hays et al., 2009; Mavletova and Couper, 2013) are rare in large-scale panel studies because of their cost and effort to implement.

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

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

2 Mode differences and previous research

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

One of the consistently-found differences between interviewer-administered and self-administered surveys relates to social desirability bias, or the tendency to present oneself in a favourable light (see DeMaio, 1984). A number of studies have found higher reports of socially undesirable behaviors, attributes, or attitudes in self-administered surveys and lower reports of socially desirable ones (for reviews, see Groves et al., 2008; Tourangeau et al., 2000; Heerwegh and Loosveldt, 2011). These findings extend to Internet surveys (see, e.g., Heerwegh, 2009; Kreuter et al., 2008). While the differences between face-to-face and telephone surveys are not as large, there is a general tendency for greater social desirability response bias on the telephone (see Holbrook et al., 2003).

Regarding the second feature of mode we explore, both face-to-face and telephone interviews involve interviewers, but may differ on the presentation of questions. Telephone is (by definition) aural, with the interviewer reading the question and response options to the respondent, who must keep this information in working memory while processing the question and formulating a response. Face-to-face surveys often involve the use of show cards, which display the response options to respondents, to minimize the cognitive burden of answering questions with several response options (see Lynn et al., 2012; Révilla, 2010). HRS does not make use of show cards, so in this respect both the face-to-face survey and telephone survey can be viewed as primarily aural modes. In contrast, the Web is a primarily visual mode, with respondents reading survey questions on the Web page. This can lead to differential response order effects, with primacy effects (in which options presented first are selected more often) occurring in visual modes and recency effects (with later options selected more frequently) occurring in aural modes (see Krosnick and Alwin, 1987; Schwarz et al., 1992; Visser et al., 2000).

3 Research questions and theoretical expectations

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

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Based on previous research, we expect more reports of depressive symptoms on the Web than in either interviewer-administered mode. Similarly, reduced social desirability biases should lead to lower reports of physical activity on the Web. However, this may be countered by response order effects (primacy on the Web), as the first option in each case indicates a higher level of activity (1 = more than once a week, 4 = hardly ever or never). In both cases, however, we expect the effect of social desirability to be stronger than that of primacy, so the overall net effect would be lower reports of physical activity on the Web.

4 Data and design

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

In addition to the biennial core interview, HRS also conducts a number of supplemental studies, mainly in the form of mail and Internet surveys that are conducted in the off-year between interview waves. The Internet survey has been ongoing since 2001 and is administered to respondents who report in their core interview that they have Internet access. The 2011 HRS Internet survey included a number of items to explore possible mode effects, repeating measures that were asked in the 2010 and 2012 core interviews. The response rate for the 2011 Internet survey conditional on having internet access, responding to the 2010 core survey and being alive in 2011 was 80%. A total of 3251 respondents who were subject to the random mode rotation completed all three surveys and comprise our analysis sample. Of these, 1583 were assigned to a telephone interview and 1668 to face-to-face in 2010. This sub-group of respondents represents 70.8% of participants in the 2011 Web survey and 14.8%/15.8% of the 2010/2012 HRS respondents. Table 1 shows the respondent characteristics in the three waves of the HRS (the two main interview waves, 2010 and 2012 and the 2011 Internet wave) and our analysis data based on a number of socio-demographic characteristics. Not surprisingly, the Internet sample (and thus our analytic sample) is more educated, younger, and comprised of higher proportions of white and coupled participants compared to the total HRS sample who were interviewed in 2010 and 2012. While this will not impact the internal validity of our results it could affect the generalizability of our findings to other populations and surveys.

The link between data collection and our analytical approach is shown in Figure 1. It can be seen that in 2010 two groups were randomly allocated to either face-to-face (Group 1) or

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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

The analysis uses a balanced panel of the respondents that took part in the 2010, 2011, and 2012 waves of the HRS. The mode variable used reflects the mode in which the interview was assigned. As noted previously, mode for the core interview was randomly assigned for panel respondents under age 80, with roughly half being assigned to telephone and half to face-to-face. Although interviewers make every attempt to complete the interview in the assigned mode, in some circumstances respondents are allowed to switch modes. Only a small proportion of respondents in our sample did not complete their interview in the assigned mode (3.1% in 2010 and 4.8% in 2012). The most common switch was from face-to-face to telephone, though some respondents also switched from telephone to face-to-face. Additionally, there are respondents who answered using the same mode in both 2010 and 2012: 155 (4.8%) answered by telephone in both 2010 and 2012 waves while 92 (2.8%) answered by face-to-face in both waves. As a sensitivity analysis all the models were rerun on the more restricted sample that includes only people who actually switched modes between 2010 and 2012. Missing data was low for the items we examine, the highest being 1.3% for the "Had a lot of energy" item.

The analysis uses Full Information Maximum Likelihood (FIML) to deal with missing data and assumes missingness at random (MAR) given the measurement model (Enders, 2010). We analysed all the models controlling for the clustering and stratification of the data as a sensitivity check (not shown). All results were consistent with those presented below.

5 Measurement models and error

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

$$y=\tau+\lambda\xi+\varepsilon \quad (1)$$

[2]As a sensitivity analysis we have also estimated the models using a wide format with correlated errors between the three measurements of the same items. The results are similar to those presented below (although for the activity scale it led to a non-positive residual covariance matrix).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where λ is the slope/loading or the strength of the relationship between the latent variable of interest, ξ, and the observed item, y. This can be considered an estimate of validity (Bollen, 1989), although it has a different meaning to that used in Classical Test Theory (Alwin, 2007; Lord and Novick, 1968). The random error, ε, can be considered an estimate of reliability (Bollen, 1989) and it can be easily calculated: $\varepsilon = 1 - \lambda^2$. Lastly, τ represents the intercept, or the threshold when the observed variable is categorical, and can be interpreted as the conditional mean or probability of the observed items when the latent variable is 0. This is usually associated with systematic error (e.g., Chen, 2008).

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

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

Using equivalence testing for estimating relative bias has become a standard procedure in cross-cultural research (e.g., Davidov et al., 2008; Van de Vijver, 2003) and it has also been implemented a number of times in the mixed-mode literature (e.g., Cernat, 2015a; Hox et al., 2015; Klausch et al., 2013). In this paper we combine the use of this procedure with the quasi-experimental design of the data collection in order to estimate the effects of modes on measurement.

Analytical approach

Using the data and the statistical method presented above we test a series of nested models to identify different types of measurement mode effects. The sequence will distinguish between random error (evaluated based on the loadings with metric equivalence) and systematic error (evaluated based on thresholds with scalar equivalence) and between modes: telephone (TEL) versus face-to-face (FTF) and interviewer versus self-administered (FTF and TEL vs. Web). From these theoretical comparisons stem the five (cumulative) models tested:

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

- **Full metric equivalence**: FTF, TEL and Web have the same loadings;
- **Interviewer scalar equivalence**: the same thresholds in FTF and TEL;
- **Full scalar equivalence**: the same thresholds in FTF, TEL and Web.

This sequence of models reflects our theoretical hypotheses regarding the impact of mode on measurement. We expect FTF and TEL to be more similar as both of them are mainly aural and involve communication with an interviewer. This would mean that it is more likely to find Interviewer equivalence than Full equivalence overall. Nevertheless some differences are expected due to higher social desirability and faster pace in TEL (Holbrook et al., 2003). On the other hand, we expect the Web to show the biggest differences from the other modes in relative systematic bias. Firstly, given it is self administered, we expect smaller social desirability effects. Secondly, it is mainly visual, which might lead to primacy effects. This leads us to expect a lack of full scalar equivalence as the systematic errors in Web are expected to be different from the interviewer modes.

It should be noted that in all these models no assumption is made about the equality of the latent variables across modes as the variances are freely estimated and the means restricted to 0 in all the modes. Thus, any learning or maturation which might appear and is not controlled for by our quasi-experimental design are expected to appear as differences in the variances of the latent variables.

To estimate the models we use Maximum Likelihood Robust estimation as implemented in Mplus 7.2. All the observed variables are considered categorical while the latent variable is modelled as continuous. As such, thresholds are calculated (the number of thresholds is one less than the number of categories) and compared across modes in order to estimate systematic error. This can be viewed either as a categorical Multi-Group Confirmatory Factor Analysis model or as an IRT model (Kankaraš and Moors, 2010; Millsap, 2012). Models are compared by using a corrected score of the $\Delta\chi^2$. This is calculated by the difference in χ^2 of two nested models. The degree of freedom of the test is the difference in degrees of freedoms between the models compared. A correction is applied to the score in order to take into account the Maximum Likelihood Robust estimation (Satorra and Bentler, 2001)[3]. The Akaike Information Criteria (AICs) are also reported. This is an indicator of relative fit based on the log-likelihood of a model that ‘penalizes’ for lack of parsimony. A smaller AIC implies a better-fitting model.

6 Results

Depression scale

We first present descriptive statistics for the nine items measuring depression (Table 2). Here we are interested in four comparisons. The first two comparisons are sensitivity checks showing the distribution of answers within the telephone and face-to-face groups by the wave in which they were answered. Here we do not expect any differences as the order was randomized. Indeed, only one of the 18 comparisons is significantly different using a Rao-

[3]See http://www.statmodel.com/chidi_shtml for explanation and an example.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Scott corrected χ^2, the 'Felt sad' item. Given this, we treat the two time points as one in subsequent analyses. The next comparison refers to our theoretical question regarding differences between telephone and face-to-face responses. Here, although we expected some amount of differences, none of the comparisons show statistically significantly results. Lastly, we treat telephone and face-to-face responses as one (Int) and compare them with Web responses. As mentioned previously, this is where we expect the biggest differences. This is indeed confirmed in the bivariate analyses as eight of the nine depression variables show significant differences between interview and Web answers.

We next model the latent depression measure with the procedure presented before using the nine dichotomous items that make up the CESD scale. The first model, Configural, assumes that the structure of the measurement model is the same across modes (e.g., no correlated errors in just one of the modes) but does not impose equality constraints on the coefficients across modes. The second model, Interviewer metric equivalence, assumes equal loadings, or reliability, across TEL and FTF. Table 3 shows that the Interviewer metric equivalence model should be selected as it does not fit significantly worse than the Configural model even if it is more restrictive (p-value of 0.85 and AIC is smaller). Similarly, the third model (Full metric equivalence), which assumes equal loadings across all three modes, fits the data well, indicating that Web does not differ in reliability compared with TEL and FTF (p-value of 0.83 and AIC is smaller). Looking at the mode effects on systematic measurement (Interviewer scalar equivalence) we find no differences between TEL and FTF (p-value of 0.72 and AIC smaller); however these two modes are systematically different from Web (p-value of 0.00 and AIC is larger for Full scalar equivalence). This indicates that the relative measurement quality is the same across modes with the exception of systematic errors between interviewer modes and Web. These results are consistent with the sensitivity analysis done using only the respondents who changed the modes in 2010–2012 and/or controlling for the mode order (not shown).

We are able to further investigate the differences indicated by these analyses. The lower part of Table 3 shows the thresholds for the two interviewer modes and those from the Web responses (from the Interviewer scalar model). Further testing shows that all the differences in thresholds are reliable with the exception of the 'Sleep' and 'Sad' variables. Constraining the thresholds of these two variables did not fit significantly worse than Interviewer scalar equivalence (Partial scalar equivalence model: $\Delta\chi^2_2=1.81$, p-value of 0.40), unlike all the other items.[4]

Because the observed variables are dichotomies (no/yes) the model estimates one threshold for each item. A large number on the threshold means that there are more people answering the first category (in this case 0 = no) while controlling for their true depression score. Differences across groups in thresholds imply relative systematic measurement differences. The results show that for all the negatively worded items that are significantly different ('Depressed', 'Effort', 'Lonely', 'Not get going'; 1 = yes = more depression) the thresholds

[4]The comparisons were made between the interviewer scalar model and models that started with full scalar equivalence but sequentially freed up thresholds. The sequence was based on the absolute difference in thresholds. Thus, the comparison started by freeing up the threshold with the biggest difference between interviewer and Web modes. If the model was significantly worse the second biggest threshold was freed, and so on.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

are lower for the Web while for all positively worded items (‘Happy’, ‘Life’ and ‘Energy’) the thresholds are higher (more no’s). This means that even after controlling for their latent score, responses in the Web mode indicated higher depression levels than those from TEL and FTF. The most plausible explanation for this pattern is higher social desirability bias in the interviewer modes. Because the scale includes both positively and negatively worded items, response order effects (primacy/recency) can be ruled out.

To make this pattern clearer we have plotted the Item Characteristic Curve (ICC, Figures 2 and 3). For example, Figure 2 plots the probability of answering “Yes” to the ‘Depression’ item, based on the latent depression score (x axis). The slope of the line is influenced by the discrimination or loading of the item. The flatter the slope, the weaker the relationship is between the item and the latent variable. The horizontal position indicates difficulty or the threshold and indicates at what levels of the latent variable the item gives information. In Figure 2, for example, saying “Yes” to the ‘Depressed’ item has a high level of discrimination (i.e., it is quite vertical), and is also is an indicator of a relatively high level of latent depression (it is positioned more to the right of the graph). What is interesting for us is how this curve is different between interviewer and Web modes. We can see that the slope of the curve is the same, due to the equal loadings, but the horizontal position is different. So, for the same level of latent depression respondents are more likely to say “Yes” to the ‘Depressed’ item on the Web than in an interviewer administered survey. To gauge the size of the effect we can choose the point on the curve that shows the highest difference across the groups. In the case of the ‘Depression’ question this is when the respondents have a score of 6 on the latent depression measure (Figure 2). At this point, based on our model, 37% of the Web responses are “Yes”, compared to 58% for interviewer administered surveys, leading to a difference of 21% in the predicted probability of answering positively.

Using the same approach, Figure 3 plots the ICC for all the questions that were significantly different between the interviewer modes and Web. It also highlights the biggest difference in predicted percentages of “Yes” for each of the item. A clear pattern obvious from this plot is that for all the negatively worded questions (first row/part A of Figure 3) the likelihood of answering “Yes” in the Web interview is higher than for the interviewer modes. On the other hand, for all the positively worded questions (second row/part B of Figure 3) this pattern reverses. This is consistent with our hypothesis regarding lower social desirability bias in the Web survey. It implies that for the same level of latent depression respondents are more likely to admit to negative affect on the Web than in interviewer modes.

In order to provide a sense of the differences between the two modes we can look at the variables with the highest and lowest significant differences (as seen in Figure 3). For example, in the case of the ‘Life’ variable for a score of 6 on the latent depression scale respondents in the interviewer-administered modes have a 67% predicted probability of a “Yes” answer compared to 34% for Web responses. We believe that this would be a substantially important difference in most applied research. At the other extreme, this difference is approximately 4% for the ‘Energy’ item.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Activity scale

The second scale we examine measures physical activity. This is based on three observed variables that ask about the frequency of different types of activities: mild, moderate and vigorous. Table 4 presents the observed distributions, first within mode by wave, and then between modes. For the sensitivity analysis it appears that two out of the six variables are significantly different using the Rao-Scott corrected χ^2: 'Moderate activity' between 2010 and 2012 response answers and 'Vigorous activity' within face-to-face. Because there seems to be no theoretical explanation or a clear pattern in the differences we combine the two waves within each mode. As with the depression scale, we find no significant differences between telephone and face-to-face but we do find differences when comparing these two with the Web answers. The patterns are in line with the response order hypothesis (with Web respondents being more likely to choose the first categories, while in the auditory modes the last ones are more likely to be chosen). The findings of the multi-group analyses below are consistent with this finding and were replicated in our sensitivity analyses when we control for mode order effects and/or restricted the sample only to people who changed mode of interview.

Next, we compare the three modes using the latent variable approach. Table 5 shows that the loadings, or reliabilities, are equal across all three modes, indicated by the fact that the second (interviewer metric equivalence) and third (full metric equivalence) models are not significantly worse than the previous ones (p-values of 0.37 and 0.12, both AICs are smaller). On the other hand, the thresholds, or relative systematic error, are the same between face-to-face and telephone (p-value of 0.98 and AIC smaller) but these two are systematically different from Web (p-value of 0.00 and AIC is larger). This implies that the level of physical activity appears to be measured systematically differently in face-to-face and telephone, on one hand, and Web, on the other. Further testing showed that only part of these thresholds is significantly different. Thus, when comparing interviewer modes with Web the third threshold for all the variables and the first threshold of "Mild activity" are significant different (Partial scalar equivalence: $\Delta\chi^2_5=4.13$, p-value of 0.53 when these are freed).

The different levels of the thresholds can be seen in the lower part of Table 5. We see that for all three variables Web respondents are less likely to choose the last category, 'Hardly ever or never', and are more likely to choose 'One to three times a month' for "Mild" at the same levels of latent physical activity. These differences are moderate to large as can be seen when we analyse the predicted probabilities of selecting a category for different scores on the latent physical activity scale. For example, looking at the predicted probability of selecting the 'Hardly ever or never' category we find a difference of approximately 9 percentage points for the "Mild" and "Vigorous" items (47% versus 38% and 57% versus 46% at a score of 3.5 and of 0.5 respectively on the latent physical activity scale for interviewer versus Web responses). The biggest difference is found for the probability of answering the same category for the "Moderate" item at a level of 1.5 on the latent physical activity variable: 81% for interviewer answers versus 31% in Web interviews. Such a pattern can be explained both by primacy/recency effects (Web respondents being more likely to choose the first categories, while in the auditory modes the last ones are more likely to be chosen), and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

higher social desirability bias when answering on the Web. While our initial expectation was that social desirability would be stronger in the interviewer modes, this does not appear to be the case. The opposite can be observed in our data as interviewer modes systematically under-report physical activity compared with Web responses. Although we cannot disentangle primacy/recency from social desirability for this scale, higher recency levels in the interviewer modes seems the most plausible theoretical explanation for the observed pattern. The absence of social desirability effects could be explained by the fact that the fitness of respondents is an observable attribute which may lower social desirability bias in interviewer modes (see Tourangeau et al., 2000, for an overview).

Impact on validity

Given these results, a natural question to ask is: how do these differences impact substantive analyses? In other words, are these differences large enough to warrant the attention of survey researchers and users? To gauge these effects we next investigate the degree to which mode changes estimates of the means and variances of the variables of interest (depression and physical activity), as well as regression coefficients (e.g., Kankaraš and Moors, 2009). We will compare three different approaches: using sum scores, using latent models but ignoring mode differences, and using latent models with correction for mode differences. The first approach might be the most typical one used in practice but it ignores both random error and mode differences. The second approach corrects for lack of reliability but ignores mode differences (it is the full scalar equivalence model from before). The last model takes into account both random error and mode effects (the partial scalar equivalence model found previously).

Table 6 shows the means[5] and variances for depression and physical activity for the three estimation approaches and the three modes. Comparing the first two columns shows how ignoring measurement error using sum scores can be biased compared to a latent model approach. Comparing the last two columns indicates to what degree ignoring the mode differences can bias means and variances. It is the last of these comparisons that is of special interest here. Overall, the differences are small. The biggest effect is for the mean depression score for the Web group, as expected. Ignoring mode differences leads to a mean which is larger than when correcting for mode effects (0.44 vs. 0.16). Although this difference is not statistically significant it might be of theoretical importance in some contexts.

Another way to estimate the impact of measurement mode effects on substantive research is to examine regression coefficients. Table 7 shows results from simple models that regress depression and physical activity in turn using six variables: if the respondent has a degree, is white, is female, is in a couple, their age and their self-rated health. Overall we see that using a sum score underestimates the regression coefficients compared to the latent models. It appears that these differences are larger than those produced by ignoring mode differences in measurement. For example, comparing the Web coefficients for the scalar equivalence

[5]In order to make the means comparable across the three approaches we have used the mean for the telephone group as the reference. Thus, we subtracted the TEL mean from the mean of the Web and FTF answers. This is similar to how means of latent variables are calculated in multi-group analysis.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

model and partial equivalence leads to small differences, none of which are significantly different.

7 Conclusions

In this paper we used a quasi-experimental design implemented in the 2010–2012 waves of the Health and Retirement Study to estimate mode effects on measurement. Using latent measurement models we compared random and systematic error on two scales: depression and physical activity. The results partially support our hypotheses regarding mode effects.

Previous literature regarding mode effects on measurement has consistently found social desirability bias as an important source of differences. This was partially replicated in our analyses. The CES-D depression scale enabled us to separate social desirability from primacy/recency effects. We show that responses collected in interviewer modes are consistently influenced by social desirability compared to Web, resulting in lower observed levels of depression even after controlling for the latent level of depression. Another possible source of mode effects we examined was primacy/recency effects. This was partially supported by our results as the Web respondents report higher levels of physical activity, consistent with higher recency effects in aural modes (i.e., telephone and face-to-face without showcards).

We have also seen that the impact of these differences on validity is relatively small. Both the means and variances of the variables and regression coefficients showed low amounts of bias due to lack of equivalence. That being said, the mean of depression was larger for the Web group in the model that ignored mode differences compared to a model that controlled for these differences. While the difference was not significant here we can imagine situations in which such differences could impact substantive conclusions. This typically happens if the reasons for the mode effects are also related to the substantive model. For example, we might be interested in the effect of self-reported BMI on diabetes. Both the dependent and independent variables can be considered sensitive and can be under-reported in interviewer modes. The mode effects on both measurements and the proportion of people in the interviewer mode(s) will determine the size of the mode effect on the relationship between BMI and diabetes. It should be also noted that the effects of the modes can bias different types of coefficients. Differences in intercepts/thresholds (as we have found here) can bias means and intercepts of substantive models while differences in loadings can bias substantive regression slopes.

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

Nonetheless, these findings have important implications for survey methodology, although they are mostly in tune with a growing body of literature on the topic. First of all, the biggest differences we found were between interviewer and self-administered modes. Our

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

hypothesized reasons, social desirability and recency/primacy, find support in our analyses. Secondly, we saw that the two scales lack equivalence in the systematic part of the measurement model between interviewer and Web modes. This implies that using a mixed-mode design may lead to lower levels of equivalence which, when combined with selection effects, could bias substantive results. Thus, a combination of improvements in design that would minimize mode measurement effects, and statistical approaches to correct for these (see Vannieuwenhuyze and Loosveldt, 2012, for overview) are advised.

Finally, in tune with other research on the topic, we caution against mixing interviewer and self-administered modes, where possible, and encourage study designs that allow for the evaluation of mode effects across a range of topics and indicators. Two such designs are the use of a reinterview in a reference mode (Schouten et al., 2013) or surveying a random subsample in the single mode of reference (Jäckle et al., 2015). The first design is a within-person one, thus making it possible to disentangle the different types of mode effects. The second approach is a between-person experimental design that allows only for the calculation of overall mode effects. As survey researchers increasingly make use of mixed-mode designs, whether in cross-sectional or panel studies, we encourage further research exploring the effects of mode changes on measurement error across a wider range of question types.

References

Alwin, DF. The margins of error: a study of reliability in survey measurement. Wiley-Blackwell; 2007.

Béland, Y., St-Pierre, M. Mode effects in the canadian community health survey: A comparison of CATI and CAPI. In: Lepkowski, JM.Tucker, C.Brick, M.De Leeuw, E.Japec, L.Lavrakas, P.Link, M., Sangster, R., editors. Advances in telephone survey methodology. John Wiley & Sons; New York: 2008. p. 297-314.

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

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

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

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

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

Chan KS, Orlando M, Ghosh-Dastidar B, Duan N, Sherbourne CD. The interview mode effect on the center for epidemiological studies depression (CES-D) scale: an item response theory analysis. Medical care. 2004; 42(3):281–289. [PubMed: 15076828]

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

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

Couper MP. The future of modes of data collection. Public Opinion Quarterly. 2011; 75(5):889–908.

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

DeMaio, T. Social desirability and survey measurement: A review. In: Turner, C., Martin, E., editors. Surveying subjective phenomena. Russell Sage Foundation; New York: 1984. p. 257-282.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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

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

Enders, CK. Applied Missing Data Analysis. 1. The Guilford Press; New York: 2010.

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

Gmel G. The effect of mode of data collection and of non-response on reported alcohol consumption: a split-sample study in switzerland. Addiction. 2000; 95(1):123–134. [PubMed: 10723837]

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

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., Tourangeau, R. Survey Methodology. 2. Wiley-Blackwell; 2008.

Hays RD, Kim S, Spritzer KL, Kaplan RM, Tally S, Feeny D, Liu H, Fryback DG. Effects of mode and order of administration on generic Health-Related quality of life scores. Value in Health. 2009; 12(6):1035–1039. [PubMed: 19473334]

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

Heerwegh D, Loosveldt G. Assessing mode effects in a national crime victimization survey using structural equation models: Social desirability bias and acquiescence. Journal of Official Statistics. 2011; 27(1):49–63.

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

Holbrook A, Green M, Krosnick J. 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. 2003; 67(1):79–125.

Hox JJ, De Leeuw ED, Zijlmans EAO. Measurement equivalence in mixed mode surveys. Frontiers in Psychology. 2015:6. [PubMed: 25759672]

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

Jäckle A, Lynn P, Burton J. Going Online with a Face-to-Face House-hold Panel: Effects of a Mixed Mode Design on Item and Unit Non-Response. Survey Research Methods. 2015; 9(1):57–70.

Kankaraš M, Moors G. Measurement Equivalence in Solidarity Attitudes in Europe Insights from a Multiple-Group Latent-Class Factor Approach. International Sociology. 2009; 24(4):557–579.

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

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

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

Krosnick JA, Alwin DF. An evaluation of a cognitive theory of response-order effects in survey measurement. Public Opinion Quarterly. 1987; 51(2):201–219.

Lord, FM., Novick, MR. Statistical theories of mental test scores. Addison-Wesley Publishing Company, Inc; 1968.

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

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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

Millsap, RE. Statistical Approaches to Measurement Invariance. 1. Routledge Academic; 2012.

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

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

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

Schouten B, van den Brakel J, Buelens B, van der Laan J, Klausch T. Disentangling mode-specific selection and measurement bias in social surveys. Social Science Research. 2013; 42(6):1555–1570. [PubMed: 24090851]

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

Steenkamp JEM, Baumgartner H. Assessing measurement invariance in Cross-National consumer research. Journal of Consumer Research. 1998; 25(1):78– 107.

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

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

Tourangeau, R., Rips, LJ., Rasinski, K. The Psychology of Survey Response. 1. Cambridge University Press; 2000.

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

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

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

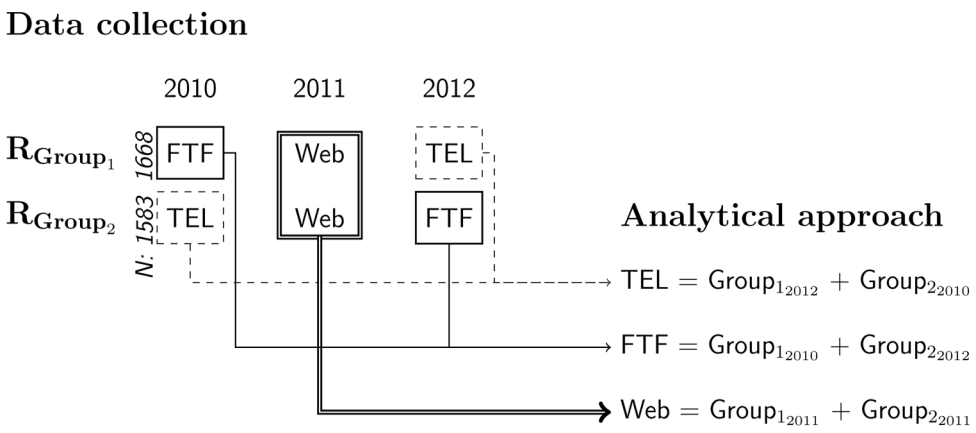

Figure 1.
The link between the quasi-experimental data collection design and analysis strategy

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

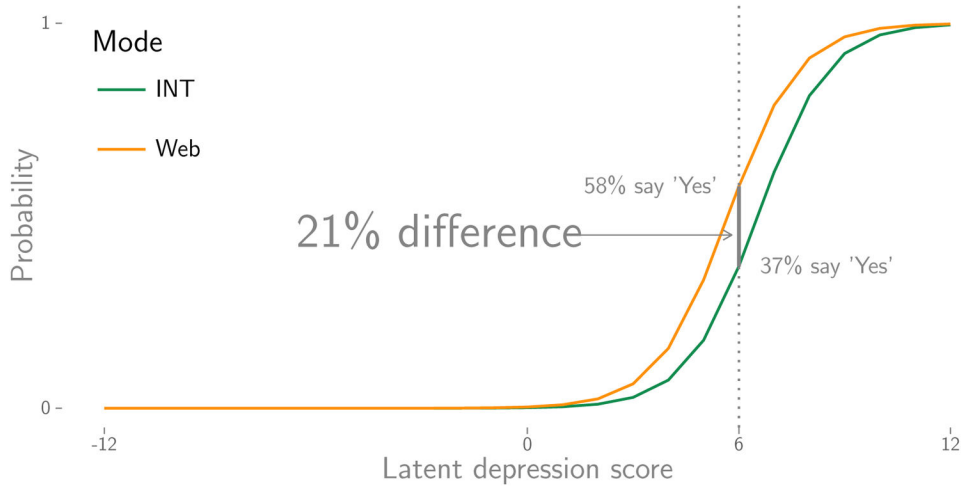

Figure 2.
Item characteristic curves for saying "Yes" to the Depression question, interviewer vs. Web.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

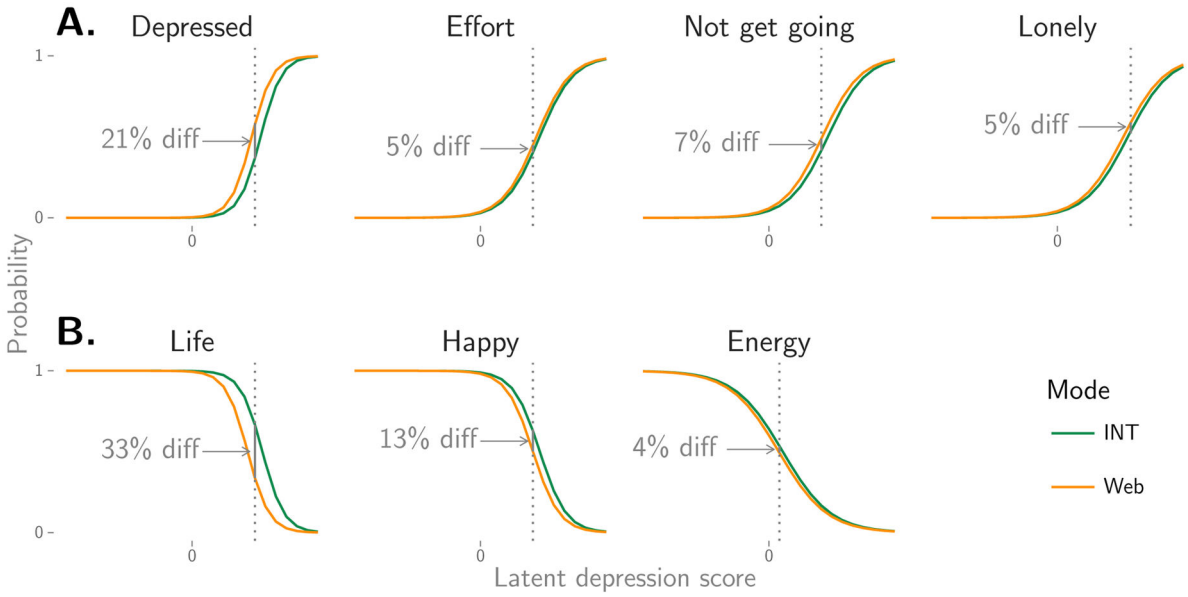

Figure 3.

Item characteristic curves for "Yes" in the significantly non-equivalent CES-D items, interviewer vs. Web. Largest difference between the groups are highlighted. First row items (**A**) are negatively worded while second row ones (**B**) are positively worded.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Composition differences between complete wave data and analysis sample (%)

		HRS10	HRS11	HRS12	Analysis
Degree	No degree	20	3	19	3
	GED	5	4	5	4
	High school diploma	47	47	47	47
	Two year college degree	6	7	6	7
	Four year college degree	13	22	13	22
	Master degree	7	12	7	13
	Professional degree	2	4	2	4
	Missing	0	0	1	0
Gender	Male	42	41	42	42
	Female	58	59	58	58
	Missing	0	0	0	0
Race	White/Caucasian	72	87	72	90
	Black or African	19	9	19	6
	Other	8	4	9	3
	Missing	0	0	0	0
In couple	No	41	27	37	25
	Yes	59	73	63	75
	Missing	0	0	0	0
	Mean age	66	65	67	65
	Base (N)	22034	4590	20556	3251

J Surv Stat Methodol. Author manuscript; available in PMC 2017 June 11.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

Sample differences in observed percentages for the depression scale. Emphasized percentages are significantly different ($p < .05$) between adjacent groups.

		TEL10	TEL12	FTF10	FTF12	TEL	FTF	Int	Web
Depressed	No	94	93	94	94	94	94	**94**	**90**
	Yes	6	7	6	6	6	6	**6**	**10**
Everything an effort	No	89	87	88	88	88	88	**88**	**86**
	Yes	11	13	12	12	12	12	**12**	**14**
Restless sleep	No	75	73	75	74	74	74	**74**	**73**
	Yes	25	27	25	26	26	26	**26**	**27**
Happy	No	11	11	10	10	11	10	**10**	**14**
	Yes	89	89	90	90	89	90	**90**	**86**
Lonely	No	90	90	90	91	90	91	**90**	**88**
	Yes	10	10	10	9	10	10	**10**	**12**
Enjoyed life	No	6	6	6	5	6	5	**6**	**12**
	Yes	94	94	94	95	94	95	**94**	**88**
Felt sad	No	88	87	**86**	**89**	87	87	87	86
	Yes	12	13	**14**	**11**	13	13	13	14
Could not get going	No	87	87	87	87	87	87	**87**	**84**
	Yes	13	13	13	13	13	13	**13**	**16**
Had a lot of energy	No	39	40	40	41	40	41	**40**	**44**
	Yes	61	60	60	59	60	59	**60**	**56**

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

Equivalence testing of the CES-D and thresholds for interviewer and Web. Highlighted model is significantly worse than the previous one.

Model	χ^2	df	$\Delta\chi^2$	p-value	AIC	Constraints
Configural	3308.315	1446			77554	Structure;
Interviewer metric equivalence	3315.851	1454	4.04	0.85	77542	Loadings: FTF & TEL;
Full metric equivalence	3357.293	1463	5.03	0.83	77531	Loadings: FTF, TEL & Web;
Interviewer scalar equivalence	3355.668	1472	6.18	0.72	77519	Loadings: FTF, TEL & Web, Thresholds: FTF & TEL;
Full scalar equivalence	3379.57	1478	99.39	0.00	77602	Loadings: FTF, TEL & Web, Thresholds: FTF, TEL & Web;
Partial scalar equivalence	3335.13	1473	1.81	0.40	77517	Loadings: FTF, TEL & Web, Part. thresholds: FTF, TEL & Web;

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

Sample differences in observed percentages of physical activities. Emphasized percentages are significantly different between adjacent groups.

		TEL10	TEL12	FTF10	FTF12	TEL	FTF	Int	Web
Mild activity	More than once a week	67	67	67	65	67	66	**66**	**69**
	Once a week	22	23	23	24	22	23	**23**	**19**
	One to three times a week	7	6	6	6	6	6	**6**	**9**
	Hardly ever or never	4	4	4	6	4	5	**4**	**4**
Moderate activity	More than once a week	**56**	**61**	59	58	59	58	**58**	**58**
	Once a week	**18**	**15**	17	17	16	17	**17**	**14**
	One to three times a week	**12**	**10**	11	11	11	11	**11**	**15**
	Hardly ever or never	**14**	**14**	13	15	14	14	**14**	**13**
Vigorous activity	More than once a week	32	32	**32**	**32**	32	32	**32**	**34**
	Once a week	11	12	**11**	**13**	12	12	**12**	**11**
	One to three times a week	11	11	**13**	**10**	11	11	**11**	**17**
	Hardly ever or never	46	44	**45**	**46**	45	45	**45**	**38**

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 5

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

Model	χ^2	df	$\Delta\chi^2$	p-value	AIC	Constraints
Configural	1251.302	153			80920	Structure;
Interviewer metric equivalence	1241.365	155	1.97	0.37	80917	Loadings: FTF & TEL;
Full metric equivalence	1224.226	157	4.17	0.12	80916	Loadings: FTF, TEL & Web;
Interviewer scalar equivalence	1226.26	166	2.45	0.98	80900	Loadings: FTF, TEL & Web, Thresholds: FTF & TEL;
Full scalar equivalence	1330.319	175	205.36	0.00	91088	Loadings: FTF, TEL & Web, Thresholds: FTF, TEL & Web;
Partial scalar equivalence	1221.83	171	4.13	0.53	80894	Loadings: FTF, TEL & Web, Part. thresholds: FTF, TEL & Web;

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 6

Mode differences in means and variances for depression and activity by estimation approach: sum scores, assuming equal measurement error (full scalar equivalence) and controlling for mode effects on measurement (partial scalar equivalence)

			Sum	Scalar eq.	Partial eq.
CESD	Mean	TEL	0.00	0.00	0.00
		FTF	−0.03	−0.05	−0.05
		Web	0.24	0.44	0.16
	Variance	TEL	3.67	14.74	15.29
		FTF	3.51	14.47	15.03
		Web	4.58	17.16	15.78
Physical activity	Mean	TEL	0.00	0.00	0.00
		FTF	0.03	0.02	0.02
		Web	−0.08	−0.05	−0.01
	Variance	TEL	6.21	1.51	1.40
		FTF	6.11	1.45	1.35
		Web	6.64	1.77	2.07

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 7

Mode differences in regression coefficients for depression and activity by estimation approach: sum scores, assuming equal measurement error (full scalar equivalence) and controlling for mode effects on measurement (partial scalar equivalence)

		Sum scale			Scalar equivalence			Partial equivalence		
		TEL	**FTF**	**Web**	**TEL**	**FTF**	**Web**	**TEL**	**FTF**	**Web**
CESD	Degree	−0.09	−0.14	0.01	−0.28	−0.37	0.00	−0.27	−0.36	−0.02
	White	0.39	0.18	0.50	0.46	0.31	0.88	0.50	0.34	0.80
	Female	0.32	0.18	0.11	0.72	0.37	0.42	0.78	0.42	0.32
	In couple	−0.39	−0.50	−0.61	−0.90	−1.09	−1.08	−0.87	−1.07	−1.12
	Age	−0.02	−0.02	−0.03	−0.05	−0.04	−0.05	−0.04	−0.03	−0.05
	Health	0.86	0.88	1.02	1.90	1.94	2.08	1.93	1.97	2.00
Activity	Degree	−0.60	−0.57	−0.57	−0.40	−0.40	−0.40	−0.39	−0.39	−0.41
	White	0.05	−0.23	−0.18	−0.01	−0.12	−0.11	−0.02	−0.12	−0.11
	Female	0.40	0.31	0.43	0.29	0.22	0.30	0.28	0.21	0.31
	In couple	−0.24	−0.20	−0.23	−0.12	−0.10	−0.14	−0.12	−0.10	−0.14
	Age	0.04	0.03	0.01	0.02	0.02	0.01	0.02	0.02	0.01
	Health	0.95	0.94	1.16	0.56	0.55	0.68	0.56	0.54	0.70