ISER Working Paper Series

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

Alexandru Cernat
Institute for Social and Economic Research
University of Essex

Mick Couper
Mary Beth Ofstedal
Institute for Social Research
University of Michigan

No. 2015-19
September 2015

INSTITUTE FOR SOCIAL
& ECONOMIC RESEARCH

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

www.iser.essex.ac.uk

Non-Technical Summary

Using multiple modes to collect data (such as face to face, telephone or the Web) is becoming a standard practice in survey agencies. While this should save costs and decrease non-response error (by including more diverse respondents) it may have detrimental effects on measurement quality. This can happen because different modes can have distinct measurement biases which, when combined with selection effects, can increase the total survey error of a mixed-mode survey relative to a single mode approach.

In this paper we use a experimental design from the Health and Retirement Study to compare the measurement quality of a number of 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 to identify measurement mode effects.

We found that 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) while religiosity shows no measurement differences between modes. Possible explanations, such as social desirability bias and primacy/recency effects, are discussed.

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

Alexandru Cernat*, Mick Couper**, Mary Beth Ofstedal**
Institute for Social and Economic Research, University of Essex

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

Abstract

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

Key words: mixed modes survey, latent measurement models, equivalence testing, Health and Retirement Study.

JEL Codes: C81, C83

Acknowledgements: We would like to thank the people that helped with this paper: Peter Lynn, Brady West, Oliver Lipps and Hayk Gyuzalyan. 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.

1 Introduction

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

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

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

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.

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

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

3 Measurement models and error

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

$$y = \tau + \lambda\xi + \epsilon \tag{1}$$

where λ is the slope/loading or the strength of the relationship between the latent variable of interest, ξ, and the observed item, y. This can be considered an estimate of reliability (Bollen, 1989), although it has a different meaning to that used in Classical Test Theory (Alwin, 2007; Lord and Novick, 1968). The random error, ϵ,

is the complement of reliability and it can be easily calculated: $\epsilon = 1 - \lambda^2$. Lastly, τ represents the intercept, or the threshold when the observed variable is categorical, and can be interpreted as the conditional mean or probability of the observed items when the latent variable is 0. This is usually associated with systematic error (e.g., Chen, 2008).

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

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

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

4 Research questions and theoretical expectations

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

Generally the HRS does not contain very sensitive questions. Many of the questions that may be subject to social desirability effects are single-item (often yes/no) questions (e.g., alcohol use, seatbelt use, smoking status), that are not amenable to our analytic approach. But both the core and Internet surveys included the Center for Epidemiologic Studies Depression Scale (CES-D) measure of psychological distress, or symptoms of depression. This consists of a series of nine yes/no items, with three items reverse-scored, which will allow us to disentangle social desirability effects from response order effects. Depression measures have been found to be subject to mode-related social desirability effects (see, e.g., Moum, 1998), although Chan et al. (2004) suggest cognitive effects related to response order may be at

work. Respondents who endorse four or more of the items are viewed as having depressive symptoms (Steffick, 2000). In addition, a three item physical activity index (frequency of mild, moderate, and vigorous exercise) was included in the Internet survey and core. Finally, we included a two item measure of religiosity (church attendance and importance of religion). As Presser and Stinson (1998) have documented, religious attendance is subject to social desirability bias associated with mode.

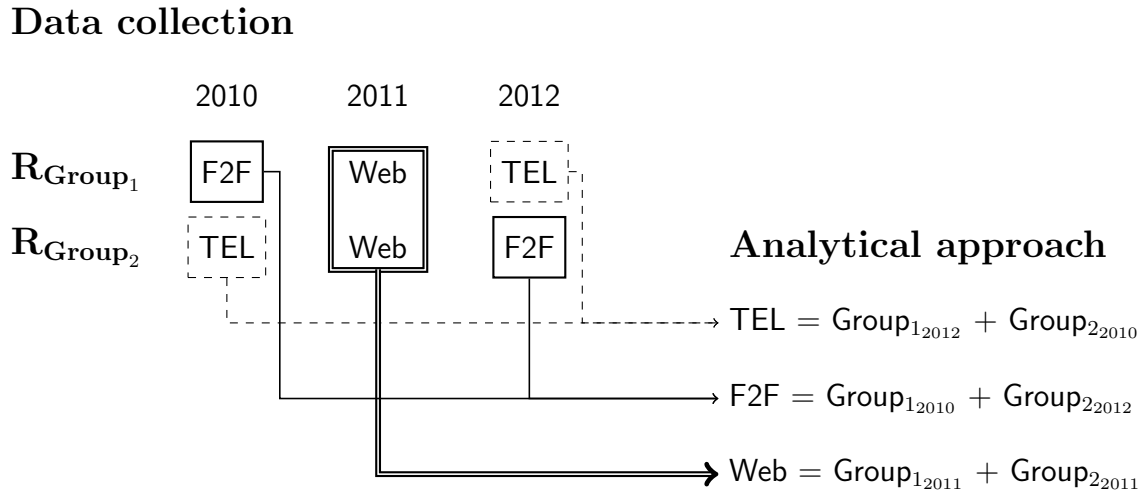
Based on the previous research, we expect more reports of depressive symptoms on the Web than in either interviewer-administered mode. Similarly, social desirability biases should lead to lower reports of physical activity on the Web. However, this may be countered by response order effects (primacy 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). Similarly, we would expect lower reports of religiosity on the Web, consistent with the social desirability hypothesis. But again, the first option for each of the two items is the high-frequency option (1 = more than once a week, 5 = not at all for religious service attendance; 1 = very important, 3 = not too important for importance of religion). In both cases, however, we expect the effect of social desirability to be stronger than that of primacy, so the overall net effect would be lower reports of physical activity and religiosity on the Web.

5 Data and design

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

In addition to the biennial Core interview, HRS also conducts a number of supplemental studies, mainly in the form of mail and Internet surveys that are conducted in the off-year between interview waves. The Internet survey has been ongoing since 2001 and is administered to respondents who report in their core interview that they have Internet access. The 2011 HRS Internet survey included a number of items to explore possible mode effects, repeating measures that were asked in the 2010 and 2012 core interviews. The response rate for the 2011 Internet survey was 81%. A total of 3251 respondents who were subject to the random mode rotation completed all three surveys and comprise our analysis sample. Of these, 1583 were assigned to a telephone interview and 1668 to face-to-face in 2010. This sub-group of respondents represents 70.8% of participants in the 2011 Web survey and 14.8%/15.8% of the 2010/2012 HRS respondents.

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



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

Data management

The analysis uses a balanced panel of the respondents that took part in the 2010, 2011, and 2012 waves of the HRS. The mode variable used reflects the mode in which the interview was assigned. As noted previously, mode for the core interview was randomly assigned for panel respondents under age 80, with roughly half being assigned to telephone and half to face-to-face. Although interviewers make every attempt to complete the interview in the assigned mode, in some circumstances respondents are allowed to switch modes. Only a small proportion of respondents in our sample did not complete their interview in the assigned mode (3.1% in 2010 and 4.8% in 2012). The most common switch was from face-to-face to telephone, though some respondents also switched from telephone to face-to-face. Additionally, there are respondents that answered using the same mode in both 2010 and 2012: 155 (4.8%) answered by telephone in both 2010 and 2012 waves while 92 (2.8%) answered by face-to-face in both waves. As a sensitivity analysis all the models

were rerun on the more restricted sample that includes only people that actually switched modes between 2010 and 2012. Missing data was low for the items we examine, the highest being 1.3% for the "Had a lot of energy" item (details can be found in the Annex). The analysis uses Full Information Maximum Likelihood (FIML) to deal with missing data and assumes missingness at random (MAR) given the measurement model (Enders, 2010).

Analytical approach

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

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

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

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

To estimate the models we use Maximum Likelihood Robust estimation as implemented in Mplus 7.2. All the observed variables are considered categorical while the latent variable is modelled as continuous. As such, thresholds are calculated (number of thresholds is one less than the number of categories) and compared across modes in order to estimate systematic error. This can be viewed either as a categorical Multi-Group Confirmatory Factor Analysis or as an IRT model (Kankaraš and Moors, 2010; Millsap, 2012). Models are compared by using a corrected score of the $\Delta\chi^2$. This is calculated by the difference in χ^2 of two nested models. The degree of freedom of the test is the difference in degrees of freedoms between the models compared. A correction is applied to the score in order to take into account the Maximum

Likelihood Robust estimation (Satorra and Bentler, 2001)[1]. The Akaike Information Criteria (AICs) are also reported. This is an indicator of relative fit based on the log-likelihood of a model that 'penalizes' for lack of parsimony. A smaller AIC implies a better fitting model.

6 Results

Depression scale

The first scale analysed using the procedure presented above is the CES-D, which estimates depressive symptoms. An underling continuous latent variable was modelled with 9 dichotomous observed items (frequencies can be found in the Annex). The first model, Configural, assumes that the structure of the measurement model is the same across modes (e.g., no correlated errors in one of the modes) but does not impose equality constraints on the coefficients across modes. The second model, Interviewer metric equivalence, assumes equal loadings, or reliability, across TEL and FTF. Table 1 shows that the Interviewer metric equivalence model should be selected as it does not fit significantly worse than the Configural model even if it more restrictive (p-value of 0.85 and AIC is smaller). Similarly, the third model, which assumes equal loadings across all three modes, fits the data well, indicating that Web does not differ in reliability compared with TEL and FTF (p-value of 0.83 and AIC is smaller). Looking at the mode effects on systematic measurement we find no differences between TEL and FTF (p-value of 0.72 and AIC smaller); however these two modes are systematically different from Web (p-value of 0.00 and AIC is larger). This indicates that the relative measurement quality is the same across modes with the exception of systematic errors between interviewer modes and Web. These results are consistent with the sensitivity analysis done using only the respondents who changed the modes in 2010-2012 and/or controlling for the mode order (not shown).

We are able to further investigate the differences indicated by these analyses. The lower part of Table 1 shows the thresholds for the two interviewer modes and those from the Web responses (from the Interviewer scalar model). Further testing has shown that all the differences in thresholds are reliable with the exception of the 'Sleep' and 'Sad' variables. When we free (i.e., allow to be different) all the thresholds with the exception of these two the model is not significantly worse than Interviewer scalar equivalence ($\Delta\chi^2_2 = 1.81$, p-value of 0.40).

Because the observed variables are dichotomies (no/yes) the model estimates one threshold for each item. A large number on the threshold means that there are more people answering the first category (in this case 0 = no) after controlling for their true depression score. Differences across groups in thresholds imply relative systematic measurement differences. The results show that for all the negatively worded items that are significantly different ('Depressed', 'Effort', 'Lonely', 'Not get going'; 1 = yes = more depression) the thresholds are lower for the Web while for all positively worded items ('Happy', 'Life' and 'Energy') the thresholds are higher (more no's). This means that even after controlling for their latent score, responses in the Web mode indicated higher depression levels than those from TEL and FTF.

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

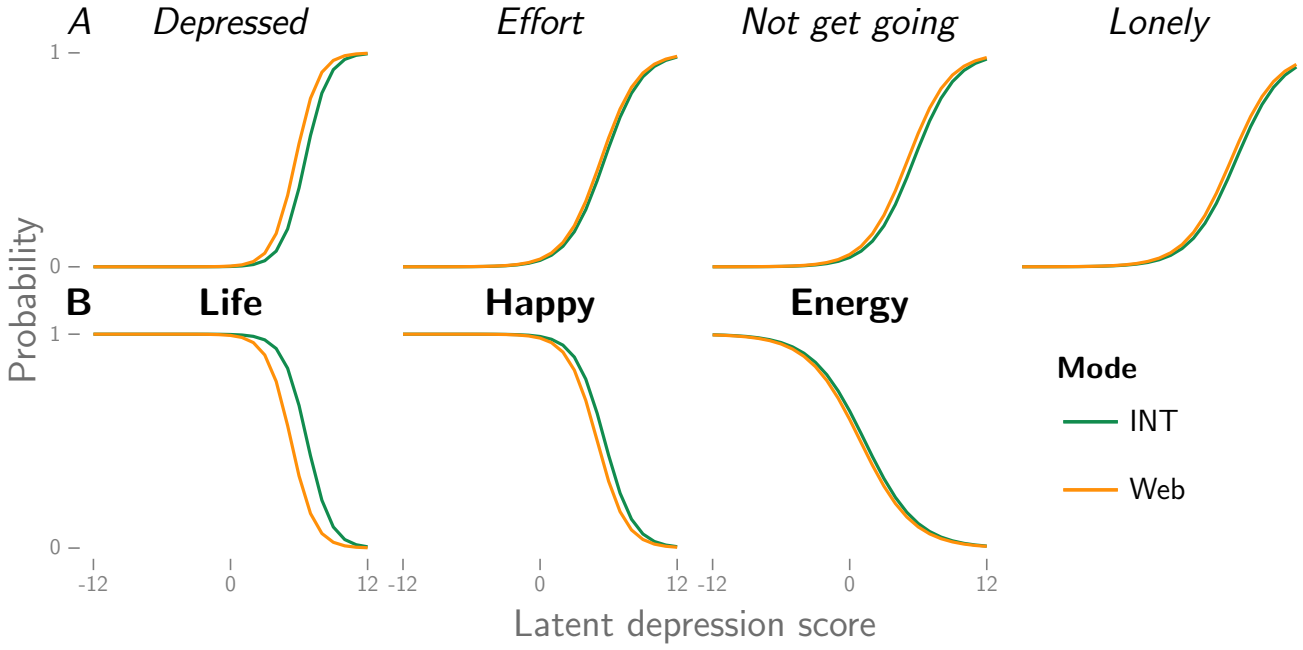
Table 1: Equivalence testing of the CES-D and thresholds for interviewer and Web.

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

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

The most plausible explanation for this pattern is higher social desirability bias in the interviewer modes. Because the scale includes both positively and negatively worded items, response order effects (primacy/recency) can be ruled out.

Figure 2: Item characteristic curves for "Yes" in the significantly non-equivalent CES-D items, interviewer vs. Web.



To make this pattern clearer we have calculated and plotted the Item Characteristic Curve (ICC) for all significant differences (Figure 2). This plots the probability

of selecting a certain category (y axis), in this case saying "Yes", based on the latent score of interest (x axis), depression. The verticality of the line is influenced by the discrimination or loading of the item. The flatter it is the less information it gives. The horizontal position indicates difficulty or the threshold and tells us at what levels of the latent variable does the item give information. In Figure 2, for example, saying "Yes" to the 'Depressed' item has a high level of discrimination, quite vertical, and is also is an indicator of a relatively high level of latent depression. What is interesting for us is how this curve is different between interviewer and Web modes. We can see that the angle of the curve is the same, due to the equal loadings, but the horizontal position is different. So, for the same level of latent depression respondents are more likely to say "Yes" to the 'Depressed' item on the Web than in a interviewer administered survey (Figure 2A). The opposite is true for positively worded items such as 'Happy' (Figure 2B). In this case one is more likely to answer "Yes" in interviewer modes given the same level of latent depression. This pattern is consistent with social desirability.

In order to provide a sense of the differences between the two types of modes we can look at the variables that have the biggest and those that have the smallest significant differences (as seen in Figure 2). Because the predicted probabilities depend on the score of the latent variable we are going to choose values on this scale that highlights the biggest mode difference for each variable/category. For example, in the case of the 'Life' variable for a score of 6 (range -12 to 12) on the latent depression scale respondents in the interviewer-administered modes have a predicted probability of 67% to say 'Yes' compared to 34% for Web responses. We believe that this would be an substantially important difference in most applied research. At the other extreme this differences is approximately 5% for the 'Effort' item (56% for interviewer surveys compared 60% for Web).

Activity scale

The second scale we analyse measures physical activity. This is based on three observed variables that ask about the frequency of different types of activities: mild, moderate and vigorous. Table 2 shows that the loadings, or reliabilities, are equal across all three modes, indicated by the fact that the second and third models are not significantly worse than the previous ones (p-values of 0.37 and 0.12, both AICs are smaller). On the other hand, the thresholds, or relative systematic error, are the same between face-to-face and telephone (p-value of 0.98 and AIC smaller) but these two are systematically different from Web (p-value of 0.00 and AIC is larger). This implies that the level of physical activity appears to be measured systematically differently in face-to-face and telephone, on one hand, and Web, on the other. Further testing showed that only part of these thresholds is significantly different. Thus, when comparing interviewer modes with Web the third threshold for all the variables and the first threshold of "Mild activity" are significant different ($\Delta\chi^2_5 = 4.13$, p-value of 0.53 when these are freed). These findings were replicated in our sensitivity analyses when we control for mode order effects and/or restricted the sample only to people that changed mode of interview.

The different levels of the thresholds can be seen in the lower part of Table 2 and their effects on the ICC's are apparent in Figure 3. We see that for all three variables Web respondents are less likely to choose the last category, 'Hardly ever

Table 2: Equivalence testing of the activity scale and thresholds for TEL, FTF and Web.

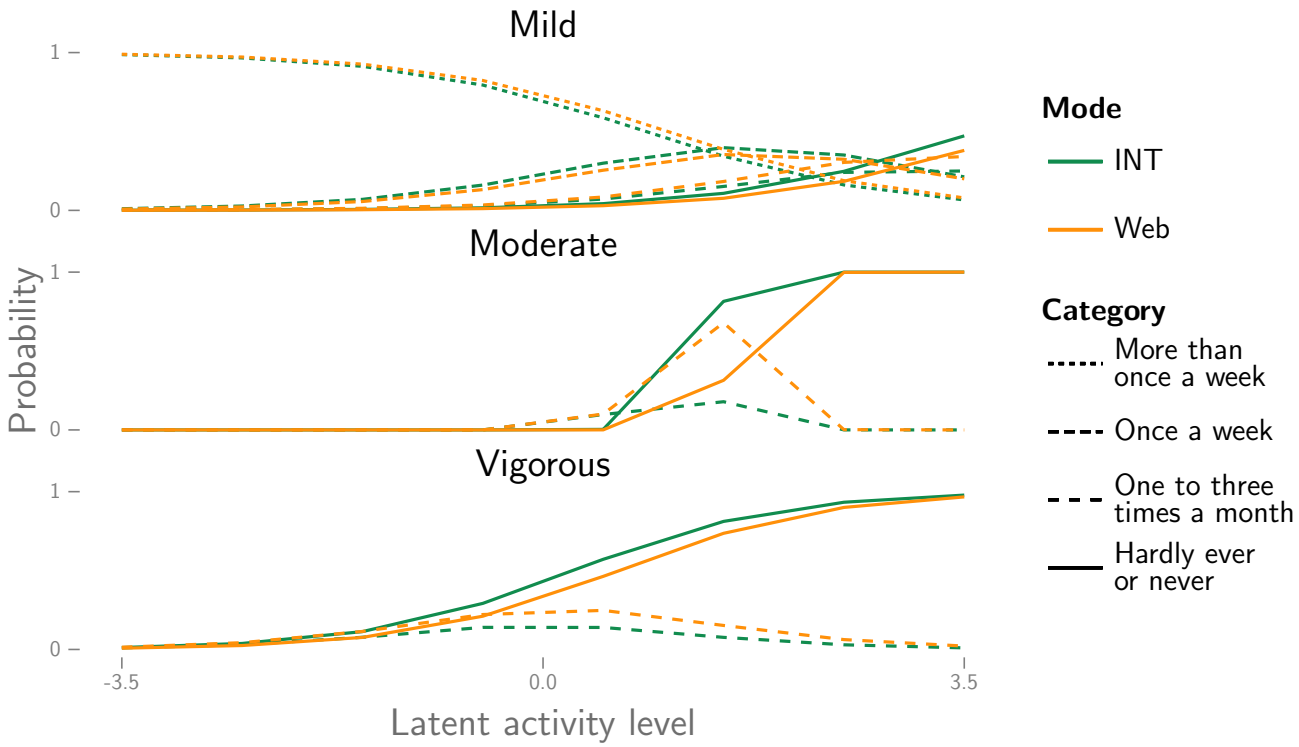
Model	χ^2	df	$\Delta\chi^2$	p-value	AIC
Configural	1251.302	153			80920
Interviewer metric equivalence	1241.365	155	1.97	0.37	80917
Full metric equivalence	1224.226	157	4.17	0.12	80916
Interviewer scalar equivalence	1226.26	166	2.45	0.98	80900
Full scalar equivalence	1330.319	175	205.36	0.00	91088

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

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

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

Figure 3: Item characteristic curves for significantly non-equivalent activity variables/categories, interviewer vs. Web.



Religiosity

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

Both the $\Delta\chi^2$ and the AIC indicate that random and systematic errors are the same across the three modes (none of the models are significantly different in Table 3). This indicates that, unlike our theoretical expectation and the two previous scales, the measurement quality of this scale is the same across modes. The sensitivity analysis, controlling for mode order and/or analysing only people who changed modes, support these conclusions as no significant difference between modes was found.

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 three scales: depression, physical activity and religiosity. The results partially support our hypotheses regarding mode effects.

Table 3: Equivalence testing of the religiosity scale and thresholds for TEL and FTF.

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

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

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

Nonetheless, these findings have important implications for survey methodology, although they are mostly in tune with a growing body of literature on the topic. First of all, the biggest differences we found were between interviewer and self-administered modes. Our hypothesised reasons, social desirability and recency/primacy, finds some support in our analyses. Secondly, we saw that two out of the three scales lack equivalence in the systematic part of the measurement model across interviewer/Web modes. This implies that using a mixed-mode design may lead to lower levels of equivalence which, when combined with selection effects, could bias substantive results. Thus, a combination of improvements in design that would minimise mode measurement effects, and statistical approaches to correct for these, such as the use of instrumental variables or of the front-door approach (Vannieuwenhuyze et al., 2014; Cernat, 2015c), are advised. The front-door approach has been recently proposed as an alternative that aims to control for causes of mode measurement effects in order to estimate selection into modes. The type of analyses carried out in this paper would be especially useful when using such a statistical approach. Finally, in tune with other research on the topic, we caution against mixing interviewer and self-administered modes, when possible, and encourage study designs that allow for the evaluation of mode effects across a range of topics and indicators.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table 4: Descriptive statistics from balanced sample of Health and Retirement Study

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

Table 4: Descriptive statistics from balanced sample of Health and Retirement Study

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