Survey Research Methods (2018)
Vol. 12, No. 1, pp. 43-58
doi:10.18148/srm/2018.v12i1.6779

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

Consequences of mid-stream mode-switching in a panel survey

Nick Allum
Department of Sociology
University of Essex
Colchester, U.K.

Frederick G. Conrad
Institute for Social Research
University of Michigan
Ann Arbor, U.S.A.

Alexander Wenz
Institute for Social and Economic Research
University of Essex
Colchester, U.K.

Face-to-face (F2F) interviews produce population estimates that are widely regarded as the "gold standard" in social research. Response rates tend to be higher with face-to-face interviews than other modes and face-to-face interviewers can exploit both spoken and visual information about the respondent's performance to help assure high quality data. However, with marginal costs per respondent much higher for F2F than online data collection, survey researchers are looking for ways to exploit these lower costs with minimum loss of data quality. In panel studies, one way of doing this is to recruit probability samples F2F and subsequently switch data collection to web mode. In this paper, we examine the effect on data quality of inviting a subsample of respondents in a probability-based panel survey to complete interviews on the web instead of F2F. We use accuracy of respondents' recall of facts and subjective states over a five-year period in the areas of health and employment as indicators of data quality with which we can compare switching and non-switching respondents. We find evidence of only small differences in recall accuracy across modes and attribute this mainly to selection effects rather than measurement effects.

Keywords: mode effects; recall; panel survey; measurement error; selection effects

1 Introduction

Face-to-face (F2F) interviews produce population estimates that are widely regarded as the "gold standard" in social research. Response rates tend to be higher with face-to-face (F2F) interviews than other modes (Hox & De Leeuw, 1994) and face-to-face interviewers can exploit both spoken and visual information about the respondent's performance to help assure high quality data (e.g. Schober, Conrad, Dijkstra, & Ongena, 2012). However, face-to-face interviews are very expensive – with marginal costs per respondent that tend to be much higher than telephone and online data collection (Groves et al., 2009; Jäckle, Lynn, & Burton, 2015). The question is whether the savings produced by these other modes outweighs any reduction in data quality. In this paper, we test the quality of data after a mode switch by comparing recall accuracy in questionnaires administered in F2F and web modes following earlier F2F data collection.

Contact information: Nick Allum, Department of Sociology, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK (E-Mail: nallum@essex.ac.uk)

Web surveys can be a cheap alternative to interviews, either F2F or by telephone. They are self-administered, eliminating interviewer costs, and there is virtually no marginal cost per case. However, one of the critical concerns in web survey research is the difficulty in garnering a probability sample. In a panel survey, it is possible to switch respondents to web mode after an initial (wave one) interview has been conducted F2F with a probability sample design, thus mitigating the problem of the lack of web-based representative sampling frames. This would appear, on the face of it, to offer an ideal solution but there is some evidence that web respondents may be more likely to take shortcuts than respondents in interviewer-administered modes (e.g Heerwegh & Loosveldt, 2008). This tendency may be further exacerbated by the experience of switching from F2F to web: by contrast to an interview, self-administration feels particularly "unsupervised" and, without the familiar experience of an interviewer to motivate them to be conscientious, web respondents may take shortcuts and minimise their effort. The result of this may be to reduce the quality of the data collected such that the gains that accrue from reducing marginal costs are offset by concomitant losses in the reliability and validity of the data thus collected. This raises the related, and more gen-

eral, issue of whether it is possible to maintain the integrity of time-series information in which there is a midstream switch of mode from F2F to web.

In this paper, we examine the viability of switching modes in this way for a major panel survey, *Understanding Society*: the UK Household Longitudinal Study (UKHLS). This survey collects data from around 40,000 households in the UK each year and began in 2009 as a fully F2F study. More recently, there have been attempts to assess the viability of moving some of the data collection to web. The UKHLS has a smaller scale supplementary survey – the Innovation Panel (IP) – where methodological improvements can be tested. This was first fielded one year before the first wave of the full UKHLS panel. This is the context for the current research, in which we examine the effect on data quality of inviting a subsample of panel members to complete interviews on the web instead of F2F. A panel survey provides rich opportunities for assessing data quality in different modes as much is known about respondents from previous waves. In the present case, we use accuracy of recall over a five-year period in the areas of health and employment as measures of data quality with which we can compare respondents switching to web with those that remain in F2F mode.

2 Background

2.1 Data quality in web compared to F2F surveys

Extant research comparing data quality and measurement error between web and F2F surveys provides mixed evidence of which data collection mode elicits higher response quality. Web surveys, on the one hand, produce significantly higher item-nonresponse rates and higher rates of "don't know" (DK) responses compared to F2F surveys (Heerwegh, 2009; Heerwegh & Loosveldt, 2008; Jäckle et al., 2015). Heerwegh (2009), for example, reports findings from a mode experiment in a student survey and shows that 8.5 percent of items are missing in web whereas only 0.02 percent are missing in F2F. Regarding "don't know" answers, Heerwegh finds that web mode produces a DK rate of around 9 percent while the average for F2F is only 6 percent. The higher item-missing and DK rates in the web mode may be due to the lack of an interviewer who is able to probe responses if survey participants are uncertain about their answer. However, item-missing rates in web data collection may be reduced if survey designers make use of interactive web survey features, such as implementing prompts for incomplete responses or using motivational statements (Conrad, Tourangeau, Couper, & Zhang, 2017; DeRouvray & Couper, 2002; Liu, Conrad, & Lee, 2016; Tourangeau, Conrad, & Couper, 2013). Interviewer-administered modes, including F2F, on the other hand, tend to induce more socially desirable responses than web (Berzelak, 2014; Chang & Krosnick, 2009, 2010; Kreuter, Presser, & Tourangeau, 2008). It seems that respondents over-report socially approved and under-report socially disapproved behaviours when communicating with an interviewer but are likely to provide more honest answers if they complete self-administered questionnaires (Groves et al., 2009).

The mode-comparison literature appears inconclusive regarding other indicators of measurement error. Although it is highly plausible that web respondents may feel sufficiently "unsupervised" during survey completion and may be more likely to adopt satisficing response strategies than F2F respondents, such as non-differentiation and acquiescence (Krosnick, 1991), the evidence is in fact rather mixed. Heerwegh and Loosveldt (2008) report that web survey participants have significantly higher levels of non-differentiation in grid questions compared to F2F respondents. Heerwegh (2009), by contrast, does not find any statistically significant difference in non-differentiation, and Berzelak (2014) has mixed findings. Similarly, Liu et al. (2016) report higher levels of acquiescence for F2F whereas Heerwegh (2009) does not find any significant difference.

Web respondents seem to select the middle category significantly more often than F2F participants (Berzelak, 2014; Heerwegh & Loosveldt, 2008) but the findings on extreme response styles are mixed: Heerwegh and Loosveldt (2008) do not find any significant difference in extreme responses between modes but both Berzelak (2014), and Liu et al. (2016) show that F2F respondents are more likely to select lower or upper extreme responses than web survey participants.

Finally, using a different type of approach, Revilla and Saris (2013) show that data quality in the European Social Survey (ESS) carried out F2F compared to that of equivalent items in the web-based LISS panel is of the same order. Here the authors use a multitrait-multimethod approach and define data quality as the strength of the relationship between latent and observed variables.

2.2 Accuracy of recall

Recall and memory have been investigated in survey experiments quite extensively, although often the purpose is either to test theories of memory (e.g. Gaskell, Wright, & O'Muircheartaigh, 2000) or to test alternate versions of a questionnaire. An overview of theories and empirical research on recall in surveys can be found in Eisenhower, Mathiowetz, and Morganstein (2004). In the present research we are examining the accuracy of respondents' recall in F2F and web questionnaires for events and facts they were asked about in an earlier F2F interview in a panel study. Because the earlier interview concerned events and states that were, essentially, contemporaneous, we treat these earlier measures as the "gold standard". While they may not be perfectly accurate, they are almost certain to be more accurate than measures taken months or years later. This allows us to evaluate

the consequences of changing modes versus reporting in a single mode over multiple waves of data collection. More specifically we ask respondents to recall their mental and objective states and behaviours in the areas of health and employment at the time when they were first interviewed that we can validate against their earlier, contemporaneous responses.

In order to provide context for our use of this retrospective response task, we briefly review here some previous studies that have assessed the validity and reliability of retrospective reports in the areas of employment and health. One example of a study on employment recall using the re-interviewing approach is Powers, Goudy, and Keith (1978). Respondents from a survey in 1964 were re-interviewed in 1974 and were asked to recall their employment situation ten years earlier. When allowing a small range of memory error around the 1964 response, 84 percent of respondents were consistent on the number of weeks employed but only 37 percent on the hours worked each week. Freedman, Thornton, Camburn, Alwin, and Young-DeMarco (1988) asked survey participants in 1985 to recall their employment status from 1980 and found that 72 percent correctly recalled whether they were in full-time, part-time, or no employment whereas 83 percent correctly identified employed vs. non-employed. Elias (1991) describes another re-interview study in which married couples who took part in a survey in 1964 were re-interviewed in 1986. Only 33 percent of women and 32 percent of men correctly recalled their employment status twenty-two years ago. A study by Mathiowetz and Duncan (1988) is an example of the record linkage approach. The authors validate respondent's recall of unemployment with company records and find that reports of the total amount of unemployment in the previous year are reasonably accurate whereas short spells of unemployment were difficult to recall.

Turning to the health field, a number of studies have assessed recall accuracy by comparing self-reports with physical examinations or medical records (e.g. Haapanen, Miilunpalo, Pasanen, Oja, & Vuori, 1997; Harlow & Linet, 1995). Much closer to our study design, previous research in the health field has involved re-interviewing the same set of respondents, comparing recalled and original responses. For example, ten Klooster, Drossaers-Bakker, Taal, and van de Laar (2007) interviewed arthritis patients about their health status and about severity of pain before the treatment. Two weeks after the treatment, they interviewed the patients again and asked them to recall their pre-treatment conditions. Comparing the concurrent and retrospective self-reports, the authors find that patients slightly over-estimate the severity of pain and poor health status. Using a similar study design, Fransson (2005) interviewed prostate cancer patients about their symptoms and quality of life prior to the treatment and then re-interviewed the patients around one year after the treatment, asking them to recall their symptoms and quality of life from one year ago. Results in the case show recall of quality of life to be quite accurate but recall of specific symptoms less so.

Research that evaluates the effect of survey mode on recall tasks is surprisingly thin and the findings point to two mechanisms through which both self-administered and interviewer modes could lead to higher quality recall. In F2F interviews, the interviewer may encourage, clarify or otherwise help respondents to recall the information required. Sudman and Bradburn (1973) found that recall about previous employment was more accurate in F2F compared to mail surveys. On the other hand, despite this potential for enhancing recall, in F2F interviews there exists a time-pressure for completing the interview that is not present in self-administered modes. To the extent that recall of past states or events takes time, web surveys, where the respondent can answer questions at her own pace, could lead to better recall. Schwarz, Strack, Hippler, and Bishop (1991) found that self-administered surveys asking respondents to recall when high profile events took place fared better than telephone interviews asking the same questions. Thus we do not have strong theoretical reasons to expect one or other mode in our study to yield more accurate recall. Furthermore, to our knowledge, there has been only one[1] study where recall accuracy has been used as an explicit indicator of data quality to compare survey modes. Morrison-Beedy, Carey, and Tu (2006) asked study participants to record their sexual behaviour on diary cards over a three-month-period. In a follow-up survey, they randomly assigned respondents to audio-computer assisted self-interview (ACASI) and self-administered paper questionnaire (SAQ) and asked them to recall their behaviour. Comparing recall and diary methods across modes, the findings are sufficiently mixed as to prevent us from making a clear prediction about differences between modes in our own research. But taken together, these studies indicate that recall tasks of the type we use in the present paper should be quite within the capacity of respondents to accomplish, but with variation in accuracy, particularly when the task requires more specific or detailed recollection. Thus, differences in recall error as a function of the mode in which respondents are asked to recall earlier states should help us evaluate the consequences of switching modes.

2.3 Mode-switching and survey context

Differences in data quality between survey modes have been examined fairly extensively in recent years (Hox, De

[1] Dillman and Tarnai (2004) also examine recall across survey modes. However, the main purpose of their study is to assess whether cognitively designed recall questions improve recall rates across modes, it is not about comparing recall accuracy between modes.

Leeuw, & Zijlmans, 2015). In many cases, researchers attribute differences between modes the affordances – properties – of the modes, e.g., primacy effects are most common when unordered response options are presented visually and recency effects are most common when the options are spoken (e.g. Schwarz et al., 1991). However, when respondents in a panel switch modes, factors besides the affordances of the modes may be responsible for differences in quality. Switching involves a change of contexts. These contexts may include environmental contexts, such as location and the presence of other people, temporal contexts, as well as internal contexts, for instance respondent's mood or fatigue during survey completion (Kelley, 2014). Whereas F2F respondents interact with a survey interviewer and are interviewed at home, web survey participants complete the survey without an interviewer being present and are able to fill in the questionnaire at any device and location of their choice. Moving from one context to another may have effects beyond those attributable to the characteristics of the target mode itself.

Particularly germane to the present research is evidence from the psychological literature that suggests that context may affect memory and recall accuracy. In one of the earliest experiments of context-dependent memory, study participants were asked to memorise a list of nonsense syllables and were tested either in a laboratory or outdoors (Smith & Guthrie, 1921). Participants who stayed in the same environment during study and test were able to recall more syllables correctly than subjects changing location. Godden and Baddeley (1975) had deep-sea divers memorise words either under water or on the shore and then switch some of them in a re-test. Recall was better when the location was the same. A recent meta-analysis (Smith & Vela, 2001) confirms these findings: Environmental context has a modest but reliable effect on memory. The more contextual elements differ between event and retrieval, the less likely are participants to retrieve the event successfully. Studies that change both environmental context and experimenter report lower levels of recall accuracy than studies changing the environment but employing the same experimenter (Smith & Vela, 2001). Since memory retrieval is cue-driven, participants are better in retrieving memories if the contextual cues at retrieval are similar to those at the experience of an event (Tulving & Thomson, 1973). If the context changes between event and retrieval, the lack of contextual cues may compromise the respondent's ability to remember the event (Smith, Glenberg, & Bjork, 1978). Environmental context has been experimentally manipulated in numerous ways, either by changing physical environments, such as different rooms (e.g. Smith et al., 1978), or keeping physical environments constant and varying elements of the environment, for example background music (e.g. Balch, Bowman, & Mohler, 1992; Smith, 1985) or odour (e.g. Cann & Ross, 1989). However, even if the physical context has changed, participants are able to mentally "reinstate" the original context of the event and generate retrieval cues if the environmental context is easy to remember (Smith, 2014). Experiment participants who were tested in unfamiliar environments but were instructed to imagine their study environments were able to recall as many words as participants who were tested in the same environment (Smith, 1979).

These findings have implications for mode-switching in panel surveys: Respondents who are interviewed in the same mode are likely to have similar contextual cues. In our case, for the F2F group there is an interviewer present in both surveys with whom the respondent interacts. Moreover, participants in our study will have experienced up to five F2F interviews, further enhancing their ability to remember their responses in a particular wave. Respondents who switch modes from F2F to web, however, lack the same kind of contextual cues in the recall interview and may find it more difficult to retrieve information from memory, which may in turn lead to lower recall accuracy among respondents who switched modes. An alternative proposition, though, consists in the following. Precisely because respondents are well-used to the F2F interview situation and amply capable of mentally imagining the context of the five previous interviews they have experienced, it may be relatively easy mentally to "reinstate" this context when confronted with the same questions in web mode. If such mental context reinstatement is successful, this would mitigate any potential attenuation in data quality arising from differences in context as well as more fundamental affordances such as visual presentation effects, social desirability, satisficing and so forth, which usually distinguish F2F and web modes.

2.4 Measurement effects and selection effects

A critical issue in studying the effect of mode, or mode-switching, in survey research is that any divergence in the distribution of responses between modes may be due in, differing degrees, to two mechanisms. Selection effects describe the situation where different types of respondents choose systematically to respond in different modes so that the true values on variables of interest differ across modes. Measurement effects on the other hand come about because of differences in measurement error between data collection modes (for a more formal treatment of this, see Vannieuwenhuyze, Loosveldt, and Molenberghs (2014). In this paper we take an approach that we hope will have useful implications for survey practice. We first want to describe the total effect of transitioning to a mixed mode design on the quality of our estimates. We do this by identifying what is called, borrowing from the medical literature, the intention-to-treat effect (ITT). This we will define as the difference in recall accuracy between those assigned to be re-interviewed in F2F mode and those assigned to complete the survey in

web mode. Our second effect of interest is the difference in recall accuracy between those completing the survey in the two modes. Again, following the lexicon of medical research we call this the effect of the treatment-on-the-treated (TOT). This is a useful quantity as it describes the magnitude of differences in estimates of recall accuracy to be expected in web mode compared to F2F. Of course, if everyone who was assigned to web complied with the invitation, these two effects would be the same. Because in our case, as in almost all realistic cases, not all respondents will want to switch modes, we need to be able to identify the "pure" measurement effects that would result for the population of respondents who agree to switching mode. This quantity is the local-average-treatment-effect (LATE), which we are able estimate by capitalising on the random assignment built in to our experiment, despite the non-compliance. More details about the way we derive these estimators is given in the next section in the paper.

Taken together, the foregoing review presents a mixture of evidence about recall accuracy, data quality and mode-switching and does not therefore suggest to us that there are unequivocal hypotheses that we should test. Rather, we simply pose the following overall research question: are there effects on recall accuracy that arise from the invitation to switch mode from F2F to web? The next section outlines in more detail how we designed our study to answer this question.

3 Data and methods

3.1 Data

The data used in this paper come from the Innovation Panel (IP) of *Understanding Society*: The UK Household Longitudinal Study (UKHLS), which is a household panel survey that has been fielded annually since 2009. The IP is a separate survey running in parallel to the main (UKHLS) study that is used to test methodological innovations and which began in 2008. In the present study we use data from Wave 1 (2008) and Wave 6 (2013) (Al Baghal et al., 2014). Wave 1 data were collected in 2008 with an issued sample of 2,670 addresses in 120 primary sampling units (PSUs) from across Britain, with 23 addresses selected per PSU, of which 1489 households provided data. All members resident in each household, including children, were asked for an interview. This yielded 2393 original sample members (OSMs) over the age of 16 in 2008, who are followed for the lifetime of the study. All of the Wave 1 interviews were carried out F2F. At Wave 6, a randomly selected subsample of approximately two thirds of households was asked, with a letter and unconditional incentive to complete the survey on the web. The remainder provided data through a F2F interview as in previous waves. For sample members assigned to web collection, a maximum of two reminders at three day

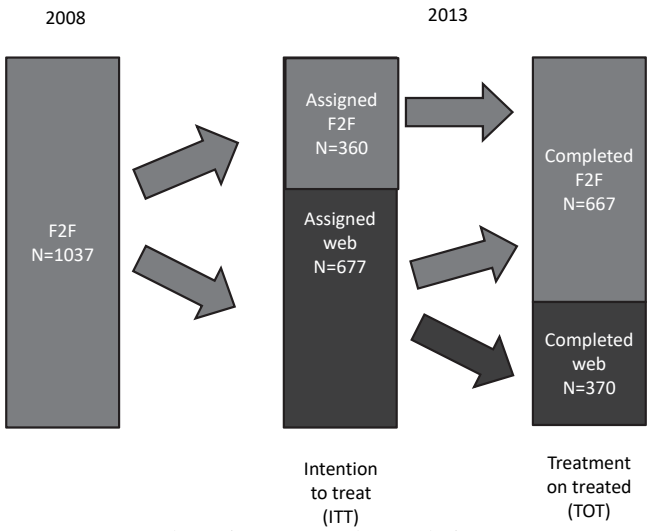

Figure 1. Sample assignment and completion

intervals were sent to individual sample members; those that had not completed the web interview within two weeks were visited by an interviewer who attempted to obtain a F2F interview. The web version of the survey remained open during the entire fieldwork period. 907 and 483 households were issued to web and F2F respectively. Our analytic sample for this paper consists of 1037 productive adult OSMs who were present, and who answered the subset of questions in which we are interested, at both Wave 1 and Wave 6. Respondents in our analytic sample of "completers" differ from the full OSM sample members, being a little older, more educated and more likely to be female, although these differences are not substantial. As discussed above, not all respondents assigned to the web condition completed web interviews. The sample design with the number of assigned and completed interviews in both modes is illustrated in Figure 1.

3.2 Study design and analysis strategy

We exploit the longitudinal data in the IP to compare data quality where mode-switching has and has not taken place. Specifically, we selected a range of subjective and objective questions, on health and employment, that were asked in the F2F interviews in 2008. We designed new versions of the same questions but which this time (the Wave 6, in 2013) asked respondents to recall their health and employment status at the time of the Wave 1 interview in 2008. By these means, we are able to assess differences in the accuracy of recall by comparing the concordance between Wave 6 and Wave 1 answers for respondents who switched to web in 2013 and those who remained in F2F.

As Figure 1 shows, there is substantial non-compliance with the request to complete by web, despite the fact that

more than 95 percent of respondents have access to the internet at home. Hence simply examining the differences in data quality comparing switchers with non-switchers conflates the effect of switching, the affordances of the mode itself and any differences, both observed and unobserved, that exist between compliers and non-compliers. In other words, mode effects on measurement are confounded with selection effects. This simple group difference estimator is the one we referred to earlier as the effect of the treatment-on-the-treated (TOT). There are a number of possibilities for dealing with this problem of confounding, and more than one quantity of potential interest that we can estimate under the most common approach for causal inference, the potential outcomes framework (Angrist & Krueger, 2001; Morgan & Winship, 2014; Rubin, 1974).

Firstly, one can attempt to adjust for potential confounders. Typically these might be a vector of sociodemographic variables. In the present case, and in general, the problem with controlling only for observables is that there may also be unobservables that correlate with both recall and mode-selection. That is to say that the selection effects may still not be "ignorable," conditioning on the observed confounders (Rosenbaum & Rubin, 1983). This problem exists whether the solution involves simple covariate adjustment or another method such as propensity score matching. For this reason, we do not adopt this strategy explicitly here, although for completeness we assess some of the observable sociodemographic predictors of compliance.

A second approach, widely used elsewhere but not often seen in the mode effects literature (cf. Vannieuwenhuyze & Loosveldt, 2013; Vannieuwenhuyze et al., 2014), is the use of instrumental variables. An instrumental variable is used in an OLS regression context when one wants to estimate the causal effect of X on Y while suspecting the presence of omitted variable bias. If X is correlated with some other causes of Y, then these effects appear in the error term, violating the assumptions of OLS. An instrumental variable, Z, is one that is correlated with X, but, by assumption, cannot have a causal effect on Y except through its effect on X. It is most often estimated via two stage least-squares (2SLS), where X is first regressed on Z and then Y is regressed on the predicted values from the first equation. Given certain assumptions, this will yield the local average treatment effect (LATE), which we introduced earlier on and which can be thought of as the effect of X on Y for the "compliers" only. In some applications this is a difficult quantity to conceptualise and may not be a useful one either. In the case of a randomised experiment with partial compliance as we have here, it is rather simply interpreted as the effect of treatment X (mode switch) for the population that would respond to the request to take part in the treatment. In the present case, this is quite a useful quantity because in a practical situation, it is impossible to compel respondents to take part in a web interview against their will or if they lack the capacity to do so.

Finally, a further approach is to examine two groups based on assignment to treatment and compare them on the outcome of interest. This estimator yields the so-called intention-to-treat effect (ITT), where the interest is in knowing the effect in the population of inviting people to take a particular treatment, knowing that not all will do so. Whilst this does, as discussed at the outset, mix selection effects with measurement effects, it is of practical use because non-compliance is the norm that we can expect, and indeed observe, in most survey situations. In our analysis we first focus on both the TOT and ITT effects on recall. At the end of the analysis we then consider the LATE estimates and examine some of the sociodemographic predictors of compliance with the request for web interviews, to provide a fuller context for our findings.

3.3 Description of variables

Our choice of variables was driven by the questions that were asked in Wave 1 of the IP in 2008. We were limited in the number of items we could add to the Wave 6 survey but we wanted a mix of subjective and objective questions in different formats that were on topics that would likely be reasonably salient for most respondents. This latter consideration was thought necessary in order that the recall task, based on a five year interval, would not be too difficult for respondents to carry out.

The topics of health and employment were both covered in the Wave 1 questionnaire and these fit our purpose well. We selected one question about employment status (Institute for Social and Economic Research, 2016), three questions on self-reported mental and physical health from the SF-12 scale (Jenkinson & Layte, 1997) and one question about long-standing illness (Office for National Statistics, 2014). The original wordings and the recall versions are shown in Table 1.

In F2F mode, showcards were used and the question and response alternatives read by the interviewer. In web mode in Wave 6, the same wordings and response alternatives as the F2F version were presented on screen. Thus, in both modes, the questions were presented visually, but in one case an interviewer also read them aloud and in the other they were fully self-administered.

3.4 Accuracy of recall

We take several approaches to assessing the accuracy of recall across experimental conditions for our five measures. Answers to the four health questions are assessed with 5-point scales or, in the case of the variable capturing limiting long term illness, a dichotomous measure. We firstly examine the differences in mean scores between contemporaneous (wave 1) and recall-based (wave 6) reports on each of

the variables and compare these differences for experimental conditions. One can think of this as a measure of net discrepancy. Secondly we use a measure of gross discrepancy, which we construct by taking the means of the absolute differences (i.e. converting negative errors to positive) between original and recall reports for each variable. For both of these discrepancy indicators a higher score will therefore signify a less accurate recall (greater discrepancies mean poorer recall accuracy). Thirdly we examine the correlations between contemporaneous and recalled reports and compare their magnitude across experimental conditions, where higher correlations would indicate greater recall accuracy.

The methods above need to be modified for our 9-category nominal measure of employment status. For this variable we use two approaches. Firstly we compare a standard measure of concordance, Cohen's kappa, computed for agreement between original and recall reports, which we can compare across experimental conditions. A higher Kappa signifies greater accuracy of recall. Kappa is a statistic that has a range of 0-1 and includes a correction for the probability of chance agreement (Cohen, 1960). The second approach is to cross-tabulate the original and recall responses and to examine whether the joint response probabilities vary by experimental condition. We do this by fitting loglinear models with interaction terms representing experimental conditions (Marascuilo & Busk, 1987).

For the purposes of testing for statistically significant differences between experimental conditions we mostly consider the health variables as a single group. We do this so as to avoid as far as possible the problem of multiple significance testing. Hence we use MANOVAs for our inferential tests for net and gross discrepancy measures and an omnibus test for equality of correlation coefficients across conditions, which we estimate using structural equation modelling (SEM).[2]

4 Results

4.1 Health variables

We begin by examining descriptive statistics for each of the variables across the two main conditions of interest, looking first at the F2F and web respondents, defined firstly by assignment status (ITT) and then by completion or actual "treatment" status (TOT). Figure 2 presents a pyramid plot showing mean scores, represented by the left- and right-most extent of each bar, on each of the four health variables at original and recall, comparing those respondents assigned to F2F at both waves with those assigned to web for Wave 6. This ITT comparison contains, as discussed earlier, a mixture of web and F2F respondents in the web condition but no web respondents in the F2F condition. The first impression upon examining this plot is that there is a remarkably close concordance between original and recall reports in both assignment groups. The three 5-point scale variables have been recoded such that high scores always indicate better health. Pain interfering with work in 2008 is slightly underestimated in 2013, and there is a marginal difference in the proportion reporting a longstanding illness at recall compared to original. These patterns look to be replicated across both assignment groups.

In Figure 3, the same comparisons are presented for the TOT groups. Here the patterns look much the same, with slightly greater disparities in recall accuracy between web and F2F groups. Overall, this first-cut look at the results does not indicate large differences arising from switching modes. It also suggests that recall is on the whole rather accurate when considering the net effects of time and mode, with no consistent trend towards under or over-estimating health status.

To drill into this a little further, we ran a paired samples t-test on the pooled samples for each of the four items where our dependent variable is the mean (net) difference between Wave 1 answers and Wave 6 recalled reports. The results of these are shown in Table 2. As well as the raw differences, a standardised effect size, Cohen's D (Borenstein, Hedges, Higgins, & Rothstein, 2009, p. 228), is shown in the final column of the table along with the correlations between Wave 1 and Wave 6. The differences are very small indeed, and not significant, for feeling depressed and general health. Cohen's d is trivial for these two items. For the remaining two, the standardised effect size is small but significant. Correlations are moderate to large for all the health items, with the largest being .55 for general health.

While the pooled net disparities are fairly small, to evaluate the "difference in differences" with respect to the ITT and TOT groups we ran two MANOVAs. The difference in disparity across all four variables between the ITT groups was not significant ($F = .48$; $p = .75$). For the contrast between TOT groups, the MANOVA was also not significant ($F = 2.25$; $p = .06$). That this test yields a smaller p-value is suggestive that, as might be expected, the TOT effect, which combines selection and measurement effects is greater than the ITT effect. Nevertheless, the more important point is that the net effects on recall from mode switching are very small.

Turning to absolute, or gross, differences, we find a similar pattern. We computed the absolute differences between scores at Wave 1 and Wave 6 for each health item for each of the four experimental groups and, again, ran a MANOVA on all four health items for both the ITT and TOT contrasts. We summarise here the results as follows. In the ITT case, there

[2] As a sensitivity test, we ran the analyses for which it was possible to use complex sample estimators (the t-tests and the regressions) that take account of the clustering in the sample design. We found no differences in the conclusions drawn and the design effects were very small. For the sake of consistency we present our inferential tests unadjusted for the clustering.

Table 1
Question wordings and response alternatives

Wave 1	Wave 6	Response alternatives
Which of these best describes your current employment situation?	Now some questions about what you were doing and how you were feeling around the first time we interviewed you. Your answers to these questions will help us improve the survey in the future. First of all, can you please tell us which of these best describes your employment situation on [Wave 1 interview date]?	Self-employed, In paid employment, Unemployed, Retired from paid work, On maternity leave, Looking after family or home, Full-time student, Long term sick or disabled, On a government training scheme, Unpaid worker in family business, Doing something else
In general, would you say your health is...	Thinking back to [Wave 1 interview date], in general would you say your health was...	Excellent, Very Good, Good, Fair, Poor
During the past four weeks, how much did pain interfere with your normal work (including both work outside the home and housework)...	Thinking back to the four weeks leading up to [Wave 1 interview date], how much did pain interfere with your normal work (including both work outside the home and housework)...	Not at all, A little bit, Moderately, Quite a bit, Extremely
Have you felt downhearted and depressed ...	And during the same period, how often did you feel downhearted and depressed...	All of the time, Most of the time, Some of the time, A little of the time, None of the time
Do you have any long-standing illness, disability or infirmity? By 'long-standing' I mean anything that has troubled you over a period of at least 12 months or that is likely to affect you over a period of at least 12 months?	And again, during the same period, that is around the [Wave 1 interview date], did you have any long-standing illness, disability or infirmity? By 'longstanding' I mean anything that had troubled you over a period of at least 12 months before that date or that you thought might affect you over the following 12 months?	Yes, No

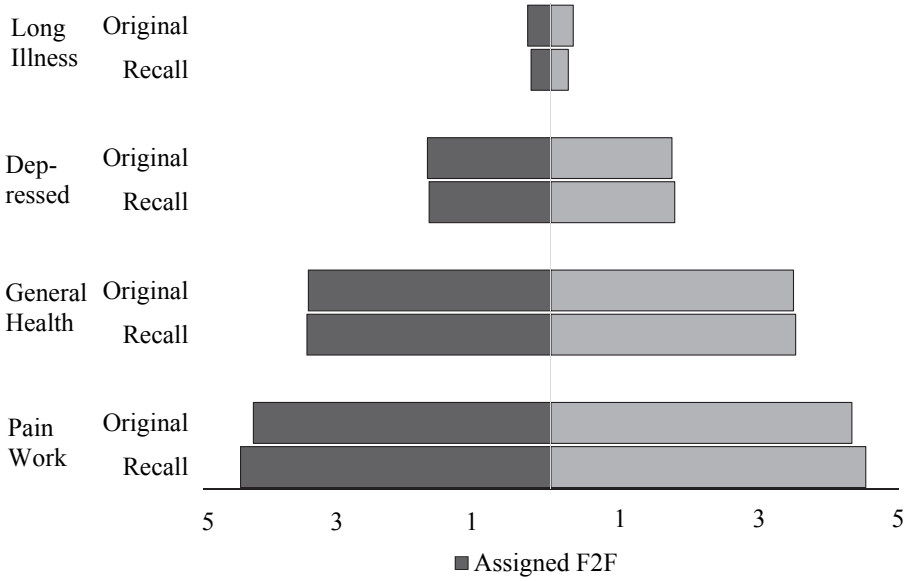

Figure 2. ITT mean scores for original and recall reports on health variables for respondents assigned to F2F and web at second Wave

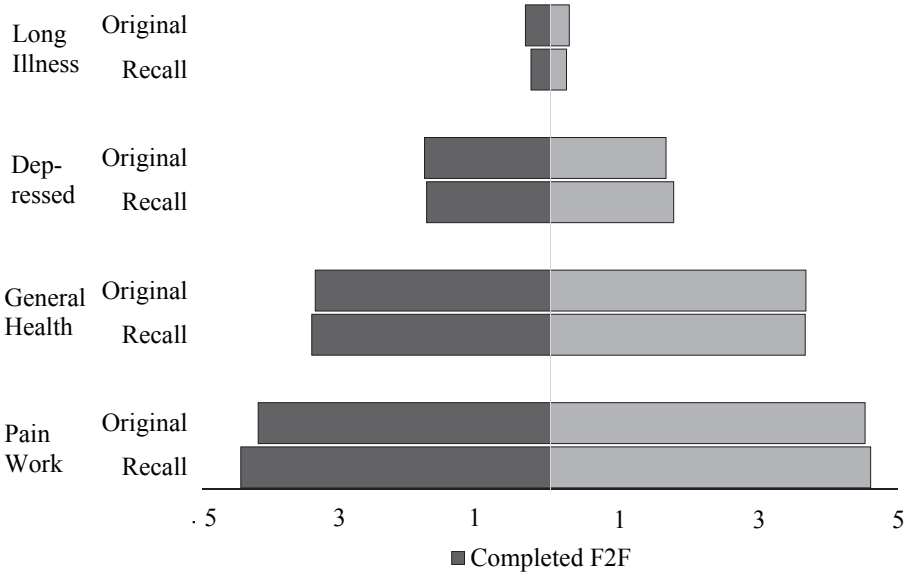

Figure 3. TOT mean scores for original and recall reports on health variables for respondents completing F2F and web at second Wave

was no significant difference in absolute disparities between treatment groups across the items (F = .37; p = .83). For the TOT contrast, the difference between web and F2F was significant (F = 4.89; p = .001).

Inspection of the individual item differences – for both net and gross disparities – indicated that one item, pain working, is driving most of the multivariate effect, and this is mainly evident in the TOT contrast only (e.g. in Table 2 Cohen's D for this variable is highest, at 0.19).

Finally, in Table 3, we break down the correlations by item between Wave 1 and Wave 6 that were shown for the pooled sample in Table 2. As can be seen, the mean correlations between original and recall reports are somewhat stronger in the F2F conditions for both ITT and TOT. We formally tested the hypothesis that correlations are equal for each item across conditions for both ITT and TOT, using SEM, and found that that, taken as a whole using the omnibus Chi^2 tests of fit for each contrast, the observed differences are greater than

Table 2
Pooled difference in means with correlations and effect sizes for original and recall reports

	$\Delta\ \bar{X}$	T-statistic	Pearson's r	Cohen's D
Depressed	−0.02	0.67	0.36	0.02
General Health	0.03	0.80	0.55	0.02
Pain Work	0.20	5.69	0.42	0.19
Long Illness	0.06	4.70	0.52	0.14

would be expected due to sampling variation alone, with both Chi^2 tests significant at <.001. We have some evidence, then, that correlations between original and recall reports tend to be a little stronger when respondents continue to be interviewed F2F.

4.2 Employment variable

Employment status is captured using a polytomous variable with eight categories. Table 4 compares the percentages selecting each category in the original Wave 1 interview (O) with the percentage recalling their employment status in 2008 in the Wave 6 interview (R). We show these percentages within each of the four experimental groups. This corresponds to a measure of net differences in original and recall reports. Strikingly, the net accuracy of recall in all of the groups is very high, with recall and original estimates matching almost perfectly. Interestingly, the probability of being employed or retired differs quite considerably across treatment groups. This is particularly true comparing TOT groups: 27 percent who completed a F2F interview reported being retired in 2008 compared to only about 15 percent of those that completed a web interview. This indicates, as expected, that TOT effects are likely confounded with selection effects. Overall, though, as with the health variables, there appears to be little difference at all in net error rates arising from switching mode from F2F to web.

While the net differences in recall error are demonstrably very small (and the overall quality of recall high) it is possible that this masks more substantial "churn" at the individual level that cancels out on average. To investigate the extent of this gross discrepancy, we examine the probability that individuals' recalled employment status matches their original report. Table 5 presents a contingency table with Wave 1 report in the columns and Wave 6 recall in the rows. Each cell contains the probability that the recalled employment status matches the originally reported status. So, for example, looking at the second column of Table 5, for those sample members that in 2008 reported being employed, the probability that in 2013 they recalled being employed is 0.9. Probabilities of matched responses are contained in the diagonal cells and are shown in boldface. As expected, the highest probabilities are on these diagonals. Gross discrepancies are smallest for the two most populous categories – employed and retired – where probabilities reach 0.9. There appears to be more slippage in the categories of unemployed and homemaker, where only around 50 percent recall their originally reported status, although sample sizes are quite small.

To test whether or not these probabilities vary systematically according to experimental treatment, we fitted two sets of loglinear models, one for each of the TOT and ITT contrasts. Each set contained original employment status, recalled employment status and the mode treatment group indicator. If mode-switching is associated with greater gross errors, we should expect that the best fitting model, short of the completely saturated one, will contain the two-way interaction of mode and recall.

The results from these models indicated that for the ITT contrast, this interaction was not necessary to reproduce the observed probabilities, whereas for the TOT the interaction with mode was needed to achieve a good fit.[3] To put it another way, this means that the small observed discrepancies between original and recall for the TOT contrasts are statistically significant.

Our final approach to examining recall accuracy for the employment variable is to compute Cohen's kappa for each treatment condition. Kappa is a measure of agreement for nominal variables that adjusts for chance agreement (Cohen, 1960). We also estimate bootstrapped 95 percent confidence intervals and present the results in Figure 4. Agreement is highest for those respondents who were assigned to and completed in F2F at both waves, at just under 0.8. Again, the TOT contrasts look more consequential, with those completing in web mode at Wave 6 having a kappa of just a little below 0.7. However, the confidence intervals all overlap, so once again this evidence is suggestive rather than definitive.

Overall, our conclusion from the analysis of the employment variable is consistent with results from the health variables. Any effects are of quite small magnitudes and suggest that such mode-switching effects as we do see are more prevalent comparing those who actually complete a web interview with those that do not, confounding selection and mode (measurement) effects to some degree. Our final analysis attempts to cast some light on this firstly by examin-

[3]Detailed results from this analysis are available on request.

Table 3
Correlation between original and recall reports by experimental condition

	ITT		TOT	
	Assigned F2F	Assigned Web	Completed F2F	Completed Web
Depressed	0.42	0.33	0.37	0.35
General Health	0.63	0.50	0.58	0.47
Pain Work	0.50	0.37	0.42	0.41
Long Illness	0.57	0.49	0.57	0.41
Mean	0.53	0.42	0.48	0.41
N	363	687	667	370
Chi^2, p	19.5, < .001		26.2, < .001	

Table 4
Employment status for original and recall reports by experimental groups

	%	Self-empl		Empl		Unempl		Retired		Home-maker		F/T student		Long-term ill		Other	
ITT	Assign F2F	8	8	57	58	2	2	19	19	6	6	2	2	5	5	1	1
	Assign Web	8	8	50	50	3	3	24	24	8	8	4	4	3	3	1	1
TOT	Comp F2F	8	7	47	48	3	5	27	27	8	5	3	3	4	4	1	1
	Assign Web	8	7	62	61	3	3	14	16	7	5	4	3	2	3	1	3

(N=1037)

Table 5
Probability of recall report matching original reported employment status

		Original							
		Self-empl	Empl	Unempl	Retired	Home-maker	F/T student	Long-term sick	Other
Recall	Self-empl	**0.70** (62)	0.03 (17)	0.03 (1)	0.01 (2)	0.00 (0)	0.00 (0)	0.00 (0)	0.10 (1)
	Employed	0.19 (17)	**0.90** (529)	0.22 (7)	0.05 (12)	0.14 (11)	0.23 (8)	0.11 (4)	0.20 (2)
	Unemployed	0.00 (0)	0.02 (9)	**0.50** (16)	0.01 (3)	0.19 (15)	0.03 (1)	0.03 (1)	0.20 (2)
	Retired	0.05 (4)	0.02 (13)	0.00 (0)	**0.91** (226)	0.11 (9)	0.01 (0)	0.06 (2)	0.10 (1)
	Homemaker	0.02 (2)	0.01 (5)	0.09 (3)	0.01 (2)	**0.53** (42)	0.00 (0)	0.03 (1)	0.00 (0)
	F/T student	0.02 (2)	0.01 (3)	0.03 (1)	0.01 (0)	0.00 (0)	**0.69** (24)	0.00 (0)	0.10 (1)
	Long-term ill	0.00 (0)	0.01 (5)	0.06 (2)	0.01 (3)	0.00 (0)	0.00 (0)	**0.78** (28)	0.00 (0)
	Other	0.02 (2)	0.01 (5)	0.06 (2)	0.00 (0)	0.03 (2)	0.06 (2)	0.00 (0)	**0.30** (3)
	Total N	89	586	32	248	79	35	36	10

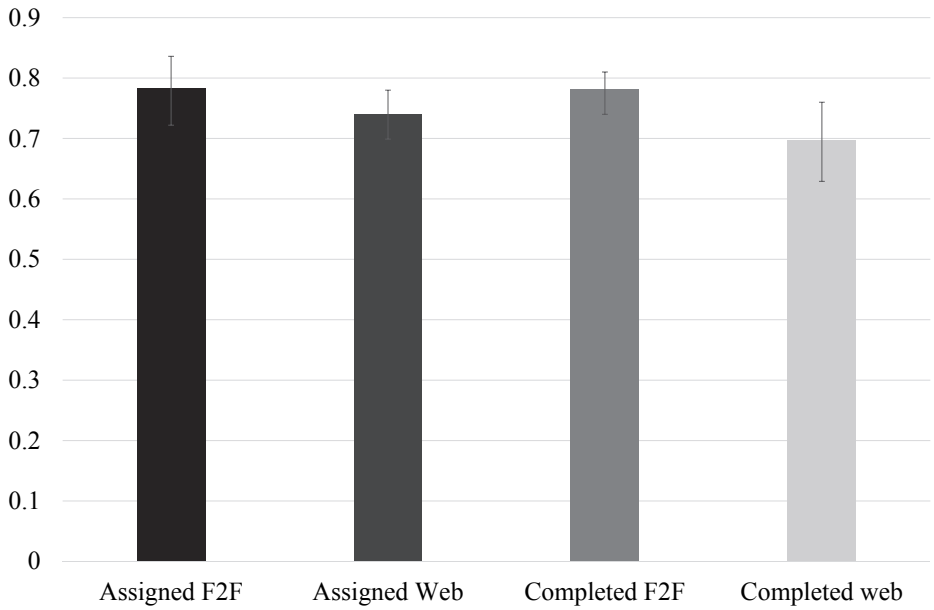

Figure 4. Cohen's kappa for employment reports by experimental condition

ing the observable differences between web and F2F respondents, and secondly fitting models that identify local average treatment effects (LATE), thus mitigating the selection problem.

4.3 Selection effects

Knowing that selection and measurement effects are confounded to some unknown degree, in this section we first present results from a logistic regression analysis to examine differences in sample composition between those respondents who comply with the request to complete an interview on the web with those that do not comply.

In so doing, we do not intend to capture all possible predictors of compliance, as many of these are likely to be unobserved, but rather give a sense of the demographic characteristics of respondents who are prone to switch mode when invited to do so. Figure 5 shows odds ratios with associated 90 percent confidence intervals from fitting a logistic regression predicting compliance, conditional on having been (randomly) allocated to web mode. As can be seen, web responders are more likely to be in professional or intermediate occupational classes than in manual classes, to be degree-educated, male and older, living in larger households. The coefficient for age squared is slightly negative, implying that the propensity for older people to comply becomes weaker for older adults. These demographic predictors are mostly in line with what we know about internet users in general (e.g. Bethlehem & Biffignandi, 2012; Couper, 2000; Couper, Kapteyn, Schonlau, & Winter, 2007; Mohorko, De Leeuw, & & Hox, 2011; Smyth, Olson, & Millar, 2014).

4.4 Measurement effects

In our final analysis, we use a different approach to disentangling selection and measurement effects for the health variables. Noting that the small effects we have seen so far are mainly found comparing those completing via web compared completing in F2F mode in Wave 6, and that there are some observable differences between these samples, we use an instrumental variable (IV) estimator to recover the Local Average Treatment Effect (LATE) for both sets of group contrasts (ITT and TOT) and for both mean and absolute discrepancies.

The LATE is also known as the complier average causal effect (CACE) and the intuition behind the approach is described earlier in the paper. Essentially, by using the experimental assignment as an instrument for actual compliance with switching to web, we can estimate what would be the effect on recall accuracy of switching to web only for those who would comply with such a request (Morgan & Winship, 2014). This isolates the causal effect on data quality of switching mode although it is only for the subset of survey respondents who are likely to switch given the choice. However, since this is in fact the main population of interest in the present application, the LATE is a useful quantity.

We computed mean and absolute Wave 1 to Wave 2 difference scores for each of the health variables and used these as the dependent variables in a set of eight IV regression models, one for each health variable and its mean and absolute difference score. We specify a dummy variable indicating whether or not the Wave 6 survey was completed on the web or not as the endogenous regressor and a dummy variable representing treatment assignment (web or F2F) as the in-

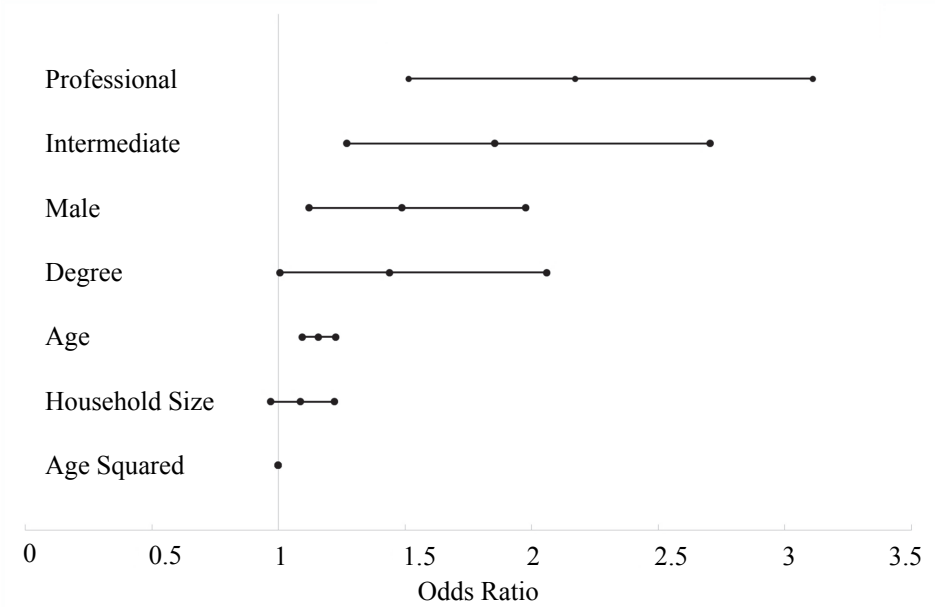

Figure 5. Completing on web conditional on assignment to web: odds ratios and 90 percent confidence intervals for selected demographic variables

strument. We do not include further covariates as we know by design that the treatment assignment is uncorrelated with the difference scores. Equations were estimated using the ivtreatreg command in Stata 14 (Cerulli, 2014).

We do not show the results here[4] because the estimated treatment effects all turn out to be very small in magnitude and uniformly non-significant. The outcome of this analysis increases our confidence in the proposition that it is selection effects, not measurement effects, associated with mode-switching that are mainly responsible for observable disparities in recall accuracy.

5 Discussion

In this paper we sought to evaluate the threats to data quality from mid-stream mode-switching from F2F to web in a panel survey. Specifically, we compared accuracy of recall of health states and employment status from five years previously, for survey respondents who had and had not been invited to switch to web mode after completing several waves of the survey with a F2F interviewer.

We find firstly that accuracy of recall is, on average, rather high for all of the experimental groups, whether based on assignment to or actual completion of a web survey for all variables. Standardised effect sizes and disparities in proportions between original and recall reports were very small for the health variables and the employment variable respectively. Gross differences in accuracy across modes were slightly larger but exhibited a similar pattern to the net effects such that the largest disparities between experimental groups were between those who completed the Wave 6 survey F2F compared to web. This was also true when we compared correlations between recall and original reports. To examine potential selection effects we compared the demographic characteristics of those respondents who agreed to complete a web survey with those who did not. The web responders (or compliers) were more likely to be male, older, with higher occupational status and educational qualifications. Additionally we estimated the local average treatment effects using an instrumental variables regression model and found that none of these LATEs were anything more than trivial in magnitude and non-significant, leading us to conclude that most of the observable disparities in recall accuracy are probably due to selection effects.

The concern in switching modes is that the context of a web survey is very different to that of the more familiar F2F interview. Since memory retrieval is cue-driven, the differing context may have impaired recall. However this seems to be only minimally true for respondents in our experiment, as recall quality is very similar after mode switch. This may be because the survey was completed at home, where F2F interviews previously took place. It may be that the rest of the survey questions in the web interview, many of them familiar to respondents, also act as environmental cues sufficient to restore accurate retrieval (Smith & Vela, 2001). Further research could examine how the presence of particular cues contribute to comparability between modes after switching.

Our overall conclusion is in one sense quite encouraging: data quality, at least that which is captured by a recall task, suffers little from switching mode in this study. On the other

[4] Available from the authors upon request.

hand, because particular kinds of persons are likely to be responsive to the request to switch, differences in observed distributions may, as we have shown, nevertheless result. Having said that, even the biggest differences seen in our analyses are rather small, so we are cautiously optimistic about the potential for maintaining the integrity of trend data in this kind of panel after a mid-stream switch, particularly if stratified by at least some of the variables that predict compliance with invitation to web.

In practical terms, for applied survey researchers, it is perhaps not the TOT effect or even the LATE that is most consequential but the ITT, which is essentially the effect of going to "mixed mode". The ITT effects are generally smaller than for TOT, although of course, the more successful any effort to switch respondents to web becomes, the more the ITT will converge to the TOT, as 100 percent compliance is approached.

6 Limitations

There are several limitations and caveats that should be mentioned. Firstly, in this study we did not consider which kind of web-enabled device was used to complete the survey. There is growing use of mobile devices for carrying out tasks that only five years ago would have usually been accomplished on a personal computer with full screen. While the results presented here combine web responses regardless of device, it may be that the relatively congenial implications of our study for mixed mode designs may be subject to revision as technology and personal communication habits develop and more people routinely use mobile devices to complete survey tasks, as there is some evidence that surveys completed on smartphones and tablets are subject to greater levels of measurement error and breakoffs (Lugtig & Toepoel, 2016). Secondly, while the use of a panel survey has permitted us to exploit the repeated waves of measurement to create a robust indicator of data quality, a relatively mature panel survey like the IP contains respondents who are already cooperative and well-practiced in answering questions. One implication of this is that the small measurement effects we observe in our results could be much amplified in a different context, such as a shorter panel or a one shot cross-sectional survey, where respondents may be less motivated to cooperate carefully with survey tasks. Thirdly, and related to the previous point, the accuracy of recall was high overall. This may have placed a "ceiling" on the size of effects from mode-switching that we could reasonably expect to see. Fourthly, although the mode of survey completion changed to web for some respondents, many of these are likely to have completed the survey at home. Thus, there may be considerable continuity in survey context and environment for these participants despite switching mode, which in turn could conceivably contribute positively to the accuracy of recall. Notwithstanding these limitations, we regard our findings as cautiously encouraging evidence for the feasibility of moving existing panel studies from traditional F2F to online as well as for the robustness of mixed and multiple mode survey research in general.

References

Al Baghal, T., Allum, N., Auspurg, K., Blake, M., Booker, C., Crossley, T., … Uhrig, J., N. S. C. Winter. (2014). *Understanding Society Innovation Panel Wave 6: Results from Methodological Experiments*. UKHLS working paper 2014-04, ISER. Colchester, University of Essex.

Angrist, J. D. & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, *15*(4), 69–85.

Balch, W. R., Bowman, K., & Mohler, L. A. (1992). Music-dependent memory in immediate and delayed word recall. *Memory & Cognition*, *20*(1), 21–28.

Berzelak, J. (2014). *Mode effects in web surveys*. Unpublished thesis. University of Ljubjana. Retrieved from http://dk.fdv.uni-lj.si/doktorska_dela/pdfs/dr_berzelak-jernej.pdf

Bethlehem, J. & Biffignandi, S. (2012). *Handbook of web surveys*. Hoboken: Wiley.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley: Chichester.

Cann, A. & Ross, D. A. (1989). Olfactory stimuli as context cues in human memory. *The American Journal of Psychology*, *102*, 91–102.

Cerulli, G. (2014). ivtreatreg: A command for fitting binary treatment models with heterogeneous response to treatment and unobservable selection. *Stata Journal*, *14*(3), 453–480.

Chang, L. & 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.

Chang, L. & Krosnick, J. A. (2010). Comparing oral interviewing with self-administered computerized Questionnaires: An experiment. *Public Opinion Quarterly*, *74*(1), 154–167.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Conrad, F. G., Tourangeau, R., Couper, M. P., & Zhang, C. (2017). Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods*, *11*(1), 45–61.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*(4), 464–494.

Couper, M. P., Kapteyn, A., Schonlau, M., & Winter, J. (2007). Noncoverage and nonresponse in an Internet survey. *Social Science Research*, *36*(1), 131–148.

DeRouvray, C. & Couper, M. P. (2002). Designing a strategy for reducing "no opinion" responses in web-based surveys. *Social Science Computer Review*, *20*(1), 3–9.

Dillman, D. A. & Tarnai, J. (2004). Mode effects of cognitively designed recall questions: A comparison of answers to telephone and mail surveys. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys* (pp. 73–93). Hoboken: Wiley.

Eisenhower, D., Mathiowetz, N. A., & Morganstein, D. (2004). Measurement Errors in Surveys. (Chap. Recall Error: Sources and Bias Reduction Techniques, pp. 125–144). John Wiley & Sons, Inc.

Elias, P. (1991). Methodological, statistical and practical issues arising from the collection and analysis of work history information by survey techniques. *Bulletin de Méthodologie Sociologique*, *31*(1), 3–31.

Fransson, P. (2005). Recall of pretreatment symptoms among men treated with radiotherapy for prostate cancer. *Acta Oncologica*, *44*(4), 355–361.

Freedman, D., Thornton, A., Camburn, D., Alwin, D., & Young-DeMarco, L. (1988). The life history calendar: A technique for collecting retrospective data. *Sociological Methodology*, *18*, 37–68.

Gaskell, G. D., Wright, D. B., & O'Muircheartaigh, C. A. (2000). Telescoping of landmark events: Implications for survey research. *The Public Opinion Quarterly*, *64*(1), 77–89.

Godden, D. R. & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*(3), 325–331.

Groves, R. M., Fowler Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology*. Hoboken, New Jersey: John Wiley & Sons.

Haapanen, N., Miilunpalo, S., Pasanen, M., Oja, P., & Vuori, I. (1997). Agreement between questionnaire data and medical records of chronic diseases in middle-aged and elderly finnish men and women. *American Journal of Epidemiology*, *145*(8), 762–769.

Harlow, S. D. & Linet, M. S. (1995). Agreement between questionnaire data and medical records. *American Journal of Epidemiology*, *141*(2), 587–592.

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

Heerwegh, D. & Loosveldt, G. (2008). Face-to-face versus web surveying in a high-internet-coverage population: Differences in response quality. *Public Opinion Quarterly*, *72*(5), 836–846.

Hox, J. J. & De Leeuw, E. D. (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys. *Quality and Quantity*, *28*(4), 329–344.

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

Institute for Social and Economic Research. (2016). *British Household Panel Survey*. https://www.iser.essex.ac.uk/bhps. Retrieved from https://www.iser.essex.ac.uk/bhps

Jäckle, A., Lynn, P., & Burton, J. (2015). Going online with a face-to-face household panel: Effects of a mixed mode design on item and unit non-response. *Survey Research Methods*, *9*(1), 57–70.

Jenkinson, C. & Layte, R. (1997). Development and testing of the UK SF-12. *Journal of Health Services Research*, *2*(1), 14–18.

Kelley, C. M. (2014). Forgetting. In T. J. Perfect & D. S. Lindsay (Eds.), *The SAGE Handbook of Applied Memory* (pp. 127–144).

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web SurveysThe Effects of Mode and Question Sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865.

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

Liu, M., Conrad, F. G., & Lee, S. (2016). Comparing acquiescent and extreme response styles in face-to-face and web surveys. *Quality & Quantity*, 1–18.

Lugtig, P. & Toepoel, V. (2016). The use of PCs, smartphones, and tablets in a probability-based panel survey: effects on survey measurement error. *Social Science Computer Review*, *34*(1), 78–94.

Marascuilo, L. A. & Busk, P. L. (1987). Loglinear models: A way to study main effects and interactions for multidimensional contingency tables with categorical data. *Journal of Counseling Psychology*, *34*(4), 443–455.

Mathiowetz, N. A. & Duncan, G. J. (1988). Out of work, out of mind: Response errors in retrospective reports of unemployment. *Journal of Business & Economic Statistics*, *6*(2), 221–229.

Mohorko, A., De Leeuw, E. D., & & Hox, J. (2011). *Internet coverage and coverage bias in countries across europe and over time: background, methods, question wording and bias tables*. Retrieved from http://www.joophox.net/papers/WebCoverage.pdf

Morgan, S. L. & Winship, C. (2014). *Counterfactuals and causal inference*. Cambridge University Press.

Morrison-Beedy, D., Carey, M. P., & Tu, X. (2006). Accuracy of audio computer-assisted self-interviewing (ACASI) and self-administered questionnaires for the assessment of sexual behavior. *AIDS and Behavior*, *10*(5), 541–552.

Office for National Statistics. (2014). *Family Resources Survey, 2007-2008. [data collection]*. Retrieved from http://dx.doi.org/10.5255/UKDA-SN-6252-2

Powers, E. A., Goudy, W. J., & Keith, P. M. (1978). Congruence between panel and recall data in longitudinal research. *Public Opinion Quarterly*, *42*(3), 380–389.

Revilla, M. A. & Saris, W. E. (2013). A comparison of the quality of questions in a face-to-face and a web survey. *International Journal of Public Opinion Research*, *25*(2), 242–253.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688.

Schober, M. F., Conrad, F. G., Dijkstra, W., & Ongena, Y. P. (2012). Disfluencies and gaze aversion in unreliable responses to survey questions. *Journal of Official Statistics*, *28*(4), 555.

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

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(5), 460–471.

Smith, S. M. (1985). Background music and context-dependent memory. *The American Journal of Psychology*, *98*(4), 591–603.

Smith, S. M. (2014). Effects of Environmental Context on Human Memory. In T. J. Perfect & D. S. Lindsay (Eds.), *The SAGE Handbook of Applied Memory* (pp. 162–182). SAGE.

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition*, *6*(4), 342–353.

Smith, S. M. & Guthrie, E. R. (1921). *General Psychology in Terms of Behavior*.

Smith, S. M. & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic bulletin & review*, *8*(2), 203–220.

Smyth, J. D., Olson, K., & Millar, M. M. (2014). Identifying predictors of survey mode preference. *Social Science Research*, *48*, 135–144.

Sudman, S. & Bradburn, N. M. (1973). Effects of time and memory factors on response in surveys. *Journal of the American Statistical Association*, *68*(344), 805–815. doi:10.2307/2284504

ten Klooster, P. M., Drossaers-Bakker, K. W., Taal, E., & van de Laar, M. A. F. J. (2007). Can we assess baseline pain and global health retrospectively? *Clinical and Experimental Rheumatology*, *25*(2), 176–181.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The Science of Web Surveys*. New York, NY: Oxford University Press.

Tulving, E. & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological review*, *80*(5), 352–373.

Vannieuwenhuyze, J. T. A. & Loosveldt, G. (2013). Evaluating relative mode effects in mixed-mode surveys: Three methods to disentangle selection and measurement effects. *Sociological Methods & Research*, *42*(1), 82–104.

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