Institute of Education

UCL

The impact of using the web in a mixed mode follow-up of a longitudinal birth cohort study: evidence from the National Child Development Study

By Matt Brown[1], Lisa Calderwood[1], Alissa Goodman[1], George B. Ploubidis[1], Joseph W. Sakshaug[2], Richard J. Silverwood[1], Joel Williams[3]

[1] Centre for Longitudinal Studies, UCL - Institute of Education, University College London. [2] Institute for Employment Research and University of Mannheim. [3] Kantar.

CENTRE FOR LONGITUDINAL STUDIES

UKRI Economic and Social Research Council

Contents

Abstract

A sequential mixed mode data collection, online to telephone, was introduced into the National Child Development Study for the first time at the study's age 55 sweep in 2013. The study included a small experiment, whereby a randomised subset of study members were allocated to a single mode, telephone-only interview, in order to test for the presence of mode effects on participation and measurement. Evidence from the experiment shows that relative to telephone-only, the offer of the web increased overall participation rates by approximately 5 percentage points (82.8% vs. 77.9%). Differences attributable to mode of interview were detected in levels of item non-response and response values for a limited number of questions. Most notably response by web (relative to telephone) was found to have increased the likelihood of non-response to questions relating to pay and other financial matters. Response by web (relative to telephone) increased the likelihood of 'less desirable' responses. For example, response by web resulted in the reporting of more units of alcohol consumed (2.4 units per week), and more negative responses to subjective questions such as self-rated health, self-rated financial status, and well-being. There was little evidence of cost-savings relative to a single-mode telephone interview. As the sequential mixed mode led to mode effects in a small number of question response rates and the value of responses, there is the potential for biases in some analyses, unless appropriate techniques are taken to correct for these.

Introduction

The availability of the web for the collection of survey data raises important questions for longitudinal studies, which need to balance potentially conflicting priorities including maximising participation, and the quality and longitudinal integrity of data, while at the same time minimising participant burden and costs. As technologies for facilitating online data collection of complex survey instruments have improved and reduced in cost, a growing number of studies, both large and small, now incorporate online data collection into their designs.

In this paper we describe and evaluate the introduction of data collection by web, via a sequential mixed mode web-to-telephone approach, that was adopted in one of Britain's renowned national birth cohort studies, the National Child Development Study (NCDS), at its age 55 sweep in 2013. This was the first birth cohort study in the UK that has used online data collection as a primary tool in one of its data collection sweeps.

NCDS is a national longitudinal study which takes as its subjects (referred to as 'cohort members') all those living in England, Scotland and Wales who were born in a single week in 1958. Cohort members have been periodically interviewed as part of the study since 1958, with the ninth and most recent follow up in 2013, when they were aged 55. Historically, the data collection mode for the 1958 birth cohort study has been face-to-face, except for the study's age 46 sweep (in 2004) which was conducted by telephone. By contrast, the 2013 survey adopted a sequential mixed-mode design (online, followed by telephone). This was the first time in the history of the survey that a mixed mode design had been adopted, and the first time that online data collection had been used.

The primary motivation for the introduction of the mixed mode was to reduce costs. However, there were other positive reasons to offer the web to participants, including optimism about response rates, driven by evidence from other studies in the UK that have shown that those aged between 50 and 65 are the most likely to respond to requests to complete a survey online (Fong and Williams 2011, Wood and Kunz 2014), and perceived limitations of the alternative interview mode available to the study at this sweep, which was by telephone.[1] The age 55 survey was also to be relatively short, a 30 minute survey as opposed to the 60 minutes or longer typical of face to face sweeps, and this was considered likely to encourage greater uptake of the online option. The cohort member was to be the sole respondent, avoiding complications arising from introducing mixed mode approaches within multiple respondent settings (see Jäckle et al 2015). Finally, the flexibility and convenience offered to study respondents was also seen as positive. Possible drawbacks included the fact that mixed mode designs may lead to so called 'mode effects', in which differences in survey responses arise simply from differences in the mode of data collection. Such mode effects can cause biases in analyses if not dealt with adequately by researchers, and by the same token create additional analytical complexity for potential users.

The introduction of the mixed mode web-to-telephone approach was an important methodological innovation in the study, and so a key priority was to build in mechanisms that would enable the effectiveness of the sequential mixed mode approach to be fully and robustly assessed. Of particular interest was to evaluate the effects of the offer of the mixed mode on overall response rates, on the final composition of the sample, and on the extent of mode effects in item-response and item-values. To this end, a random subgroup of around 1 in 8 members of the NCDS issued sample were allocated directly to the telephone as a single mode, rather than to the sequential mixed mode. This embedded experiment enabled an evaluation of how the sequential mixed mode approach compared to the counterfactual of a telephone-only study design on these dimensions.

In this paper we provide a first assessment of the success of the sequential mixed mode approach adopted in NCDS, based on the results from the embedded experiment, and on contextual evidence regarding the cost of the exercise. Overall, the experience appears to have been positive in several dimensions, most notably on the overall study response rate, which we show to have increased by approximately 5 percentage points as a result of the sequential mixed mode approach. However, the sequential mixed mode also led to mode effects in a small number of question response rates and the value of responses, which has the potential to lead to biases in some analyses, unless simple and appropriate techniques are taken to correct for these. The evidence on costs is not conclusive, however our best guess is that in this instance cost savings were either zero or very modest, relative to the option of a single mode telephone-only approach.

The structure of the paper is as follows. First, we provide a literature review and outline our research questions. Next we describe the NCDS study and provide further details of the sequential mixed mode design that was introduced. Then we describe the experiment that was embedded within the mixed mode design and set out the methodology used to evaluate the embedded experiment. Next, we provide evidence on the balance of the samples in the mixed mode and telephone only arms of the random assignment. Then we provide information about response and non-response rates to the survey in both groups, and set out the characteristics of responders of different types. We then set out the main findings from the randomised experiment, namely the effect of assignment to mixed mode, and the effects of response by web on survey response, item response and item values. Then we explain what

[1] Telephone was the only alternative option due to budget constraints.

lessons can be drawn about cost savings in this setting. Finally, we conclude by setting out potential lessons for NCDS and for other studies considering introducing the web into its design and explain further work underway in relation to understanding the impact of the introduction of the mixed mode approach.

Literature review and research questions

In response to decreasing response rates and rising costs associated with implementing large-scale face-to-face surveys, longitudinal and cross-sectional surveys are making increasing use of mixed-mode data collection strategies, especially strategies which involve the web (Jäckle, Gaia, Benzeval 2017; De Leeuw 2018). Long running longitudinal surveys, such as the UK Understanding Society, the UK Next Steps Cohort study, the US Panel Study of Income Dynamics, and the US Health and Retirement Study have begun (or are planning) to use web in a mixed-mode design. The potential for improved response rates, reduced risk of nonresponse bias, and cost savings are key motivations behind the shift towards mixing modes. However, evidence on the actual impact of introducing web as part of a mixed-mode design within longitudinal surveys is limited. Our main focus lies with sequential mixed-mode designs (as opposed to concurrent designs), which deploy multiple modes of data collection in a specified order. Sequential mixed-mode designs can be cost-effective when they start with a less expensive mode, such as web, and then switch to a more expensive mode, such as telephone or face-to-face, for nonresponse follow up (Hochstim 1967; Siemiatycki 1979; McHorney, Kosinski, and Ware 1994; McMorris, Petrie, Catalano, Fleming, Haggerty et al. 2009; Wagner, Arrieta, Guyer, and Ofstedal 2014).

Although web surveys tend to produce lower response rates than other modes (Manfreda et al. 2008; Daikeler, Bosnjak, Manfreda 2019), there is evidence that combining web with an interviewer-administered mode in a sequential mixed-mode design can produce higher response rates relative to an otherwise equivalent design without web (Greene, Speizer, and Wiitala 2008; Kappelhof 2015; Elliott et al. 2009; Sakshaug, Cernat, and Raghunathan 2019). However, this result has not been replicated in the few experiments implemented within large-scale longitudinal studies. For instance, an experimental mode design study implemented in the Understanding Society Innovation Panel wave 5 found that sample members assigned to a sequential mixed-mode design with web followed by face-to-face interviews participated at a lower rate compared to sample members assigned to the unimode face-to-face design (Jäckle, Lynn, and Burton 2015). This effect dissipated in subsequent waves as there were no differences in attrition when the same experiment was implemented in waves 6 and 7 of the Innovation Panel (Bianchi, Biffignandi, and Lynn 2017). The authors also reported only minimal differences in respondent composition between the two mode designs. The same experiment, implemented in wave 8 of the main Understanding Society, also showed no increase in attrition rates between the sequential web-face-to-face and face-to-face designs (Carpenter and Burton 2018). It should be noted sample members assigned to the mixed-mode group were offered higher incentives than those in the unimode group. Gaia (2017) reports that this strategy was indeed effective in increasing participation in the mixed-mode group to a level that was comparable to that of the unimode group.

While introducing a sequential mixed-mode design with web in a longitudinal study may not significantly improve response rates, there is suggestive evidence that it can yield significant cost savings. Cost savings can arise through high web take-up rates which preclude inter-

viewer involvement. Bianchi et al. (2017) report an increasing share of respondents who participated via web in the mixed-mode treatment design of waves 5 (42.7%), 6 (55.6%), and 7 (57.5%) of the Understanding Society Innovation Panel. Given that these households did not require an interviewer in the mixed-mode group, the estimated cost savings were around 10%, 14%, and 23% in the respective waves after accounting for incentive costs. In the Next Steps age 25 survey, a sequential mixed-mode design with web followed by telephone and face-to-face was implemented which resulted in about 61% of respondents participating via web (Calderwood 2016). The use of additional incentives for web completion boosted web response rates and led to cost savings of around 25,000-30,000 GBP due to fewer cases being issued to face-to-face interviews.

Despite their purported cost savings, mixed-mode strategies that involve the web are susceptible to data quality issues, including item non-response and differential measurement errors. Item non-response tends to be higher in web and other self-administered modes than in interviewer-administered modes (de Leeuw 2005; Heerwegh 2009; Greene, Speizer, and Wiitala 2008; Heerwegh and Loosveldt 2008; Hope et al. 2014; Scott et al. 2011). Consequently, adding web to an otherwise interviewer-administered design has the potential to increase item non-response. Jäckle, Lynn, and Burton (2015) report significantly higher rates of “don’t know” and refusals under the mixed-mode design in wave 5 of the Innovation Panel. Across 1055 items, the average item nonresponse rate was about 65 percent higher in the mixed-mode group than in the face-to-face group.

On the measurement error side, it is well-known that mode can influence the way in which people answer survey questions (De Leeuw 2005). That is, respondents might give different answers to the same question depending on which mode they are interviewed in (Jäckle, Roberts, and Lynn 2010). For example, it is well-known that respondents interviewed in self-administered modes provide fewer socially desirable responses (Greene, Speizer, and Wiitala 2008; Kreuter, Presser, and Tourangeau 2010; Laaksonen and Heiskanen 2014; Heerwegh 2009) and contribute less positivity bias (Ye, Fulton, and Tourangeau 2011; Hope et al. 2014) compared to respondents interviewed in interviewer-administered modes. Survey modes are also susceptible to presentation effects. For instance, self-administered modes tend to elicit more primacy effects due to their visual presentation while interviewer-administered modes are more prone to recency effects due to the aural administration of the questionnaire items (Krosnick and Alwin 1987). Moreover, complex questions involving detailed instructions or definitions may be challenging to administer in self-administered modes due to lack of interviewer support. All of these mode-related measurement effects could lead to potential differences in response distributions between mixed-mode and unimode designs. Only Jäckle (2016) has explored this issue experimentally, finding differences for about 3% of items collected under the web-face-to-face and face-to-face treatment groups in Understanding Society.

The paucity of experimental evidence on the effects of switching to a mixed-mode design involving web in a longitudinal study represents a clear research gap in the literature. Several open questions remain regarding the impact of using web in conjunction with interviewer-administered modes. For example, the reviewed literature suggests that introducing a web-face-to-face design in a longitudinal study has a negative (or no effect) on participation and item nonresponse rates, and may come as a shock to panel members who have grown accustomed to being interviewed face-to-face. Whether this finding is consistent across other studies, involving different mode combinations (e.g. web-telephone), is unclear. Further, it is unclear the extent to which a mixed-mode design with web yields different response distributions compared to a unimode design without web. While introducing web to an interviewer-administered survey is expected to reduce social desirability bias, we do not know whether

such reductions translate in a sequential mixed-mode design, or are washed away by other factors that influence responses under a given mode design (e.g. selection error, aural vs. visual presentation). Lastly, it is important to obtain a better understanding of the cost implications of switching to a mixed-mode design. Web take-up rates are important in this regard but the initial wave of implementation also carries significant start-up costs related to programming, infrastructure, pretesting, etc. Although these start-up costs might be expected to dissipate in subsequent waves of a longitudinal study, they could offset any initial cost savings that may result from issuing fewer cases to face-to-face interviews.

To shed further light on these issues, we make use of a mixed-mode design experiment implemented in the NCDS. We use these data to address the following research questions:

1. Does introducing a sequential web-telephone design yield a similar response rate, relative to a telephone-only design? Is the likelihood of participation in the web-telephone design uniform across respondent subgroups?
2. Does the web-telephone design result in higher rates of item nonresponse relative to the telephone-only design?
3. To what extent are survey items affected by introducing a web-telephone design relative to the telephone-only design? Is there evidence that social desirability is reduced under the mixed-mode design?
4. What is the impact of introducing the web-telephone design on survey costs? How high is the web take-up rate under this design? Are cost savings evident relative to the telephone-only design?

The NCDS study and the sequential mixed mode in its age 55 sweep

The NCDS is an ongoing multidisciplinary cohort study of all babies born in Great Britain in a single week in 1958. The initial birth survey was conducted by midwives in hospitals across Great Britain, and participants have subsequently been followed up at 7, 11, 16, 23, 33, 42, 44, 46, 50 and 55 years of age. The initial sample of 17,415 individuals was augmented during childhood by immigrants into Great Britain, with a resulting total sample of 18,558. The initial sample at birth contained 98.1% of all babies born in Great Britain in the study week, and even after more than 5 decades, retention remains very high, with 9,137 study members taking part at the age 55 sweep in 2013.

From the original focus on the circumstances and outcomes of birth, the study broadened in scope to map all aspects of health, education and social development as the cohort passed through childhood and adolescence, while in adult life the information collected has covered education and training, labour market activity, housing, family formation, income, health and well-being. At a biomedical sweep at age 44, physical measurements and biological samples were also taken.

Most previous sweeps of the study have been conducted face-to-face by interviewers in the cohort members' homes. One exception to this was the age 46 sweep in 2004, which was a telephone interview. The introduction of the telephone interview was part of a broader strategy for the cohort, according to which a less expensive mode of interview – namely the telephone – would be used for alternate study sweeps. Thus whilst 'major' study sweeps would retain the face-to-face interview mode and would cover all substantive areas related to age

and life stage, 'minor' study sweeps, which would be shorter, focused mainly on updating life event histories and a subset of topic areas of particular relevance to age and life stage. This pattern was followed for the age 46 sweep (a 'minor' sweep, by telephone), and age 50 (a 'major' sweep, conducted face-to- face), and initial plans for the age 55 sweep were to conduct a telephone interview, according to the original strategy. However, following a request from the funder to further cut costs, and in consideration of some potential positives benefits to the study of adopting an online approach, the study opted for a sequential web-to-telephone design.

The sequential mixed mode design was implemented as follows: initially, all cohort members were asked to complete the questionnaire online. Non-responders (after six weeks and three letters/emails) were contacted by telephone (where possible) and asked to do a telephone interview instead. When designing the survey every effort was made to ensure equivalence between the web and CATI instruments, drawing extensively on the Unimode design principles set out by Dillman and colleagues (2009). In addition, the great majority of the content of the survey was factual in nature, and such questions are generally acknowledged as being less prone to mode effects (e.g. Lozar et al., 2002; Schonlau et al., 2003)[2]. The resulting questionnaire was approximately 30 minutes long and covered: household composition, housing, economic activity, qualifications, help and care provided to parents and grandchildren, earnings, income and housing wealth, retirement plans and pensions, self-reported health and health conditions, smoking, drinking, well-being, and the updating of job and partnership event histories.

While the large majority of cohort member were allocated to the mixed mode protocol, a subset of cohort members were randomly allocated to a single mode, telephone-only protocol, which we describe below.

Measuring mode effects: the mixed-mode experiment, and methodology for its evaluation

The use of different data collection modes both over time (in a longitudinal study), and within a study sweep using a sequential mixed mode approach, as adopted by NCDS at age 55, introduces the possibility of mode effects in the data. Mode effects are present if responses to items differ across individuals, or within individuals over time, solely due to the mode in which the response is given, rather than due to differences in the underlying constructs which the questionnaires or other data collection instruments are designed to capture. Where different mediums are adopted at different data collection sweeps, mode effects may occur longitudinally, i.e. they may affect the measurement of change over time for the same individuals. While these may be important, we do not consider evidence for these in this paper. The sequential mixed-mode design adopted for the age 55 sweep of NCDS also raises the possibility of mode effects occurring cross-sectionally, within a given sweep. Here mode effects may affect the measurement of differences between individuals at a point in time, since by design individuals within the same survey provide their responses by different modes, with each individual choosing just one of the possible response modes offered. If

[2] A full account of the design decisions taken when developing the web and telephone questionnaires is provided by Brown (2016) (https://cls.ucl.ac.uk/wp-content/uploads/2017/04/CLS-WP-20162.pdf).

such mode effects exist and are not dealt with through appropriate statistical methods, they may lead to biases in analyses.

Typically within sequential mixed mode settings, it is difficult to detect and to correct for such mode effects robustly, simply because it may be impossible to distinguish differences in responses between individuals that are due to *measurement,* and those that are due to *selection,* the latter occurring because individuals choose (or 'select into') their mode of response. For example, it has been demonstrated in this and other sequential mixed mode data collection contexts, that those choosing to answer by web are wealthier, more likely to live with a partner and, unsurprisingly, more likely to be regular web users (Wood and Kunz, 2014) than those opting to respond in subsequent modes. In NCDS such observable differences are captured in data drawn from the long history of participants' prior participation in the study and thus can potentially be controlled for, for example through multivariable regressions or using propensity score matching. However, even with a rich set of prior controls there may yet be further unobserved differences between individuals choosing between response modes that make attributing differences in responses to either measurement or selection difficult.

In order to investigate fully the extent of mode effects within the age 55 sweep, and furthermore to be able to assess the extent to which any mode effects detected may bias analyses and to enable users to robustly correct for any such biases, a random subset of cohort members was therefore allocated to a telephone-only data collection protocol. We refer to this as the 'mixed mode experiment'. In the main stage of data collection at the age 55 sweep of NCDS a total of 11,553 addresses were issued to the fieldwork agency. The experiment included 10,586 cohort members with UK telephone numbers (thus the experiment excluded 967 cohort members, most of whom were emigrants from the UK who were allocated to a web-only protocol, and others with no phone number). Among those included in the experiment, 1,476 cohort members – or around 1 in 8 – were allocated to the telephone-only group, with the remaining 9,110 allocated to the mixed mode group. The proportion allocated to the telephone-only group was chosen as an adequate sample size to be able to detect any substantial mode effects that might bias inferences.

Embedding an experiment of this type allows us to disentangle measurement from selection effects, and specifically to estimate the overall impact on outcomes of interest of employing a mixed-mode data collection approach compared to a telephone interview data collection approach (called the 'intention to treat' effect within the evaluation literature). We can also, under some credible assumptions (spelled out further below), estimate the impact of responding by web for the subgroup that completed the survey online (here referred to as the 'complier average causal effect'), which is the treatment of most interest for the understanding of mode effects.

Methodological approach to the evaluation of mode effects

One simple methodological framework we can use to obtain estimates of these mode effects is to conceptualise the mixed mode experiment as a **randomised experiment with one-way non-compliance** (Imbens and Rubin, 2015). Following the notation of Imbens and Rubin, the treatment of principal interest here, W_i, is defined as a response by web (as compared to a response by telephone) for cohort member i = 1, …, N. We are interested in the causal effect of W_i on outcomes of interest, denoted Y_i. Outcomes of interest for us include overall participation in the study sweep, as well as the individual item responses in the sweep, and item values. Additionally, the embedded mode experiment gives us an instrument, Z_i, which is defined as the random assignment to the mixed mode treatment group, rather than the telephone-only group. This instrument is a priori known to have a causal effect

on W_i (since only those in the mixed mode group can respond by web). For convenience, these definitions are summarised in Table 1, while the structure of the experiment and the outcomes to be evaluated are illustrated further in Figure 1.

In any experimental setting, non-compliance refers to a situation where individuals are randomised into a treatment group, and some choose to comply with that treatment, while others do not.[3] In our context those who are randomly assigned to the mixed mode group may respond by web or by telephone to the survey. Compliers can be thought of as those who respond by web, while those choosing the telephone option within the mixed mode group are thought of as the non-compliers within the group. Among the control group (telephone-only), the only response option is by telephone (and hence non-compliance in this context is one-way only[4]). Since compliance (the web) is voluntary rather than enforced, the group of compliers is a selected group, and despite the experimental setting, additional assumptions are required to estimate the effect of the treatment on outcomes of interest.

Within this framework two different treatment effects on outcomes Y_i can be clearly uncovered:

The first is a simple Intention to Treat (ITT) estimator, giving the difference in the expected value of Y in the treatment vs control group.

$ITT_Y = E(Y(Z=1) – Y(Z=0))$

This ITT effect is not the effect of the treatment, but only assignment. As such this captures the effects of the offer of the mixed mode, relative to a counterfactual of a telephone-only survey design, on outcomes of interest. The ITT effect is given by the difference in mean outcomes between the treatment and control arms of our experiment, and gives us helpful information about how the offer of the mixed mode in age 55 NCDS affected responses compared to the survey having been conducted as a fully telephone-only study sweep. The main drawback of this ITT analysis is that it does not answer questions about causal effects of web response itself, only about causal effects of the overall assignment to the mixed mode group.

One important insight given by Imbens and Ruben is that this overall ITT effect (ITT_Y) can be understood as consisting of two parts, the overall treatment effect on the compliers (who responded by web) and the non-compliers (who responded by telephone), weighted by their population proportions. Assuming that the treatment effect on non-compliers is zero, or in other words that the offer of the web has no impact on the responses of cohort members who chose to answer by telephone, this then provides us with the second estimator of interest in this paper.

The second estimator that we derive is known as the Complier Average Causal Effect (CACE)[5], and is the one which is of most relevance to us for capturing mode effects since it gives us the causal effect of answering by web relative to answering by telephone. This estimator is defined as the ratio of the ITT effect to the population proportion of compliers, as follows:

[3] The example given by Imbens and Rubin is a Vitamin A supplement trial, where not all of those randomised into a group receiving Vitamin A supplements chose to take Vitamin A, and it is desired to know the effect of Vitamin A supplementation on one or more outcomes.
[4] Non-compliance is 'one way' in our context because while those in the mixed mode group could opt for the telephone as opposed to the web, web response was simply not available to responders in the telephone-only group.
[5] This same effect is also sometimes also referred to as a Local Average Treatment Effect, or LATE.

CACE = ITT_Y/ ITT_W, where ITT_W = E(W(Z = 1) – W(Z = 0)) = E(W(1))

This effect, under a plausible assumption (known as an 'exclusion restriction'), provides us with an estimate of the key mode effects of interest, where these are defined precisely as the average causal effect of responding by web (compared to by telephone) for those who responded by web. The exclusion restriction applied here is (as stated above) the assumption that among the non-compliers in the mixed mode group – i.e. those in the mixed mode group who chose to answer by telephone – the offer of the web did not affect the answers that they gave in the telephone survey. While we cannot directly test this assumption, this seems at face value to be a reasonable assumption to make.

Given the experimental nature of these estimates, the estimators for the ITT_Y are derived by taking simple differences between sample means; the ITT_W is directly measured as the population proportion of compliers within the group Z = 1 (i.e. the proportion who responded by web among all responders within the mixed mode group), while CACE estimators are then subsequently derived by straightforward division of the ITT_Y by this scalar.

It should be noted that non-response to the survey, and to individual items ("NR" in Figure 1), are further options for survey respondents among both the mixed mode and telephone-only groups, which means that the experimental comparison between the two groups may be violated when examining item responses and individual survey items. We examined the extent of such selection into response by comparing the unit responders and non-responders in each experimental group. Subsequently, we controlled for a number of pre-treatment characteristics in analyses examining item responses and individual survey items. Due to the necessity to control for pre-treatment characteristics, ITT effects were in practice estimated using regression approaches and CACEs using instrumental variable regression approaches.

Furthermore, since pre-treatment characteristics were not observed for all cohort members included in the mixed mode experiment, conducting a complete case analysis would have reduced the analysis sample size and potentially introduced bias. We therefore utilised a multiple imputation (MI) approach in which the imputation model included the experimental group assignment and the pre-treatment characteristics to be used as control variables[6], with further variables known to be predictive of unit non-response[7] included as auxiliary variables. We used MI by chained equations and generated 5 imputed datasets. Forthcoming analyses will increase to 20 the number of imputed datasets and further refine the imputation model for the later analyses, though results are not expected to change markedly.

All analyses were conducted using Stata version 15 (StataCorp, College Station, TX).

[6] Whether participated at 50, whether email provided to study, computer at home, use home computer >2 per week, personal access to internet, computer skills excellent or good, sex, white British, homeowner, in work, professional/managerial occupation, weekly net pay, living with partner, has degree, fair or poor health, regular smoker, problematic drinker, CASP-12 Quality of Life Score, animal naming test score, word list recall score, delayed word list recall score, letter cancellation speed score, letter cancellation accuracy score

[7] Social class at birth, region of residence at birth, Rutter behavioural score at age 7, social environment at age 7, region of residence at age 11, special educational needs at age 11, English ability at age 11 and general ability test score at age 11.

Balance of experimental groups

The balance of the experimental groups was examined by comparing a set of pre-treatment characteristics taken from the age 50 NCDS sweep. Table 2 shows the means of these pre-treatment characteristics by randomly assigned experimental group status. There was no evidence of differences between the mixed mode and telephone-only groups. A further joint test of the differences between the two groups using a multivariable probit model also confirmed that there was no evidence that the groups were unbalanced.

Survey response, compliance status, and non-response

For the total issued sample for the mainstage fieldwork of 11,553, Table 3 presents the issued sample and response rates, according to whether cohort members were part of the experiment – and hence allocated to either the mixed mode or telephone-only groups – or were not included in the experiment. Assignment to the mixed mode generated a higher response rate (82.8%) compared to the telephone-only survey protocol (77.8%). The table also shows response rates for the group that (for various reasons) were not included in the experiment, and who were allocated to a web-only protocol. This shows a low response rate particularly among cohort members who are not known to have emigrated but for whom no valid UK telephone number was held.

Table 4 provides information about compliance to the treatment by showing the overall proportion of those allocated to the mixed mode who chose to respond by web. In total, 5,612 of the 9,110 cohort members allocated to mixed-mode data collection completed the questionnaire online (61.6%). Although a high online response rate had been anticipated, this result exceeded initial expectations set at the design stage by more than ten percentage points. A further 1,935 completed interviews by telephone, representing an additional 21.2% of the mixed mode group. Web-responders represent 74.4% of all responders in the mixed mode group – this proportion represents the population proportion of compliers, or $ITT_{W.}$

Amongst those allocated to the mixed mode approach, the pre-treatment characteristics of those who chose to complete via the web (so-called ‘compliers’ with the treatment) were markedly different than those who participated via telephone – thus confirming that there is strong selection on observable characteristics into response by web (Table 5). There was no evidence of gender difference in response by web, but there was evidence of a difference in all other characteristics considered. Those who chose to complete by web were more likely to have participated at age 50, to have provided the study with an email address, to have had a computer at home, to have been regular home computer users, to have had internet access and to rate their computer skills positively. In terms of socio-economic characteristics, web completers were more likely to have been in work at age 50, to have been in professional/managerial occupations and to have had higher net earnings. Web completers were also more highly qualified and more likely to have lived with a partner at age 50. Web completers reported better health at age 50 and higher levels of well-being, and were also less likely to smoke or have alcohol problems. Web completers also achieved higher scores in each of the four cognitive assessments administered at age 50.

A final set of comparisons in this section compares the pre-treatment profiles of responders and non-responders within the mixed mode and telephone-only groups (Table 6). Despite the fact that response to the survey was higher among the mixed mode than the telephone-only group, there were very few differences between the profiles of the responders to the two groups. There was some evidence that respondents in the telephone-only group were very slightly more likely to have participated in the age 50 sweep than respondents in the mixed mode group (95% vs. 94%) and that mixed mode respondents were more likely to have been defined as problematic drinkers at age 50 (18% vs. 15%), though in the context of multiple testing these results should not be over-interpreted. Taken as a whole, these results suggest that the two data collection strategies did not affect the overall balance of observable characteristics of those who chose to participate. This is important since it gives some face validity to the experimental comparisons of item responses and item values between responders across the two treatment arms.

Evidence of mode effects on participation and on item response and values

Here we present estimates of the two main treatment effects outlined previously: the experimental comparison between the mixed mode and telephone-only groups, the so called intention-to-treat (ITT) estimator, which provides an estimate of the offer of the mixed mode, and the complier average causal effect (CACE), which provides an estimate of the effect of responding specifically by web, compared to by telephone.

Survey participation

As already noted, assignment to the mixed mode group generated a higher response rate (82.8%) compared to the telephone-only group (77.8%). This equates to a 5 percentage point ITT effect (95% confidence interval [CI] 2.7, 7.3) (Table 7). This was little changed upon control for pre-treatment characteristics to 5.2 (95% CI 3.2, 7.2), as would be expected given the balance of these covariates between the experimental groups.

Item non-response

Comparisons of some item non-response rates are shown below in Table 8. For some items there was strong evidence of differences between the mixed mode group and the telephone-only group, but for many items no difference was apparent. Where differences occurred it was typically the case that higher item non-response rates were found amongst the mixed mode group (as shown by the ITT estimates), and we thus conclude that web response (relative to telephone) causes higher non-response in those items (as shown by the CACE estimates). Most of the largest differences related to variables where a numeric value had to be entered – value of home, amount left to pay off on mortgage and gross weekly pay. These questions are all fairly sensitive meaning that one might have expected that the anonymity of the web would have led to a lower item non-response rate in the mixed mode group than in the interviewer administered telephone-only group. However, accurately answering these questions would also require a considerable degree of cognitive effort and so it seems that telephone interviewers may have encouraged telephone respondents to provide an answer. Somewhat surprisingly, the item non-response rate on the ‘type of employer provided pension’ question was higher amongst the telephone-only group than the mixed mode group. This is a complex question and one might have assumed that the presence of an interviewer

who could potentially have provided clarification to any queries raised by the respondent would have led to a lower item non-response rate amongst the telephone-only group, but this clearly was not the case.

Mode effects in item values

Table 9 shows the extent of any mode differences for a subset of variables which are likely to be widely used by analysts. In terms of factual socio-economic variables there was little evidence of mode effects. There was no difference between the mixed mode group and the telephone-only group in terms of reporting being in work, having a professional/managerial occupation, number of hours worked per week, weekly pay (gross and net) or housing wealth. The small number of socio-economic variables where a mode difference was found tended to be based on questions which employed a rating scale and it was typically the case that telephone-only respondents responded more positively than those in the mixed mode group. For example, telephone-only respondents gave a more positive rating of their current financial situation and reported a lower likelihood of working at the age of 66.

This positivity bias amongst the telephone-only group was also apparent with relation to health variables. Telephone-only respondents reported better self-rated general health on average, were less likely to give responses which lead them to be classified as disabled or having a long-standing illness, and tended to report a lower number of specific health problems. Mixed mode respondents were more likely to report suffering from a subset of specific health problems, namely back problems, hearing problems and depression. There were no significant differences between the mixed mode and telephone-only group in terms of the prevalence of reporting being a regular smoker, nor in the reported frequency of consuming alcohol, but the average reported number of units of alcohol consumed in the last 7 days was higher amongst the mixed mode group. The telephone-only group also reported better well-being across all five well-being variables and higher levels of different leisure activities.

In terms of voting, the sole difference was that the mixed mode group were more likely to have reported voting Liberal Democrat in the 2010 election.

Evidence on whether introducing the web brought cost savings

As noted in the introduction to this paper, one motivation for introducing the mixed mode design with the age 55 sweep of NCDS was due to the potential for cost savings. The potential for cost savings was promising, given that NCDS is a study of individual cohort members, rather than whole families (unlike Understanding Society, for example (Buck and McFall 2012)), which meant that the achievement of cost savings was dependent only on whether cohort members as individuals chose to respond online, rather than on the decision also taken by other family members.

While we did not collect any experimental data that would robustly enable us to assess whether adopting the sequential mixed mode web-to-telephone approach relative to a short telephone-only survey saved costs, indicative evidence from tenders received to a competitive tendering exercise for fieldwork suggests that there was little difference in cost between tenders for a web-to-telephone approach, and telephone-only. In part this was due to high initial development costs for the web – with a particularly resource-intensive activity being the development of interactive event histories for the web survey.

While cost savings were perhaps not realised by adopting the web on this occasion, web costs would most likely be reduced in the future, now that initial development of web interfaces has taken place. Moreover, cost savings could be higher in future now that high web take-up is known (tenders were costed assuming 40% take-up rate, relative to over 60% achieved). Cost savings might also be higher for longer interviews.

Conclusions

The higher response rate obtained for the sample allocated to the mixed mode data collection protocol is a clear benefit to the adoption of the mixed mode approach. The higher than expected participation by web, at over 60%, showed that the majority of the cohort was willing (and indeed preferred) to complete the questionnaire online rather than respond by telephone, with an overall gain in response of around 5 percentage points from the adoption of the mixed mode relative to the telephone-only approach a very positive one for the study.

Online and telephone data quality appears to be comparable in most cases – as evidenced by the lack of detected mode effects in most variables collected in the study sweep. However, there were some clear exceptions to this – most notably in self-assessed ratings of financial status and health, self-reports of a number of health conditions and indices of well-being, and in item non-response to financial variables such as income/earnings and wealth, all of which are heavily used survey items of central interest to many users of the study.

Where mode effects have been detected, the inclusion of an experimental element to the survey has enabled us both to inform researchers of exactly which variables present mode effects, as well as allowing the potential for robust methodologies to correct for these.

What implications do these findings have for future data collections in NCDS or other birth cohort and other studies? In order to employ the mixed mode approach judiciously, we mainly restricted data collection to more factual topics. Telephone and online data are not usually compatible if the question asks for a judgment on a topic (especially non-salient ones), asks about non-norm behaviours or attitudes, or uses a multi-point response scale. Our findings of significant mode effects in the key subjective questions that were included, and in other questions where social desirability bias was likely to be an issue, confirm this judgement as correct. Perhaps less expected were the differences in item non-response in financial variables, which suggests the web, and mixed mode approaches which include the web, may also not be ideal for these items.

Going forward in the design of future sweeps, these limitations suggest that sequential mixed mode web-to-telephone should not become the primary mode of approach, especially for the study's major sweeps where collection of financial variables, for example, is a central part of the study. In any case, the use of the telephone seems likely to prove increasingly challenging as part of any survey design. However, the successful take-up by web, and its apparent superior data quality to phone, especially for some question types – for example, where social desirability appears to be an issue – suggests potential continued use of the web as part of a multi-mode approach to future sweeps, perhaps as a supplement to face to face interviewing.

References

Bianchi, A., Biffignandi, S., and Lynn, P. (2017). Web-Face-to-Face Mixed-Mode Design in a Longitudinal Survey: Effects on Participation Rates, Sample Composition, and Costs. *Journal of Official Statistics*, 33(2), 385–408.

Buck, N., and McFall, S. (2012). Understanding Society: Design Overview. *Longitudinal and Life Course Studies*, 3(1), 5-17.

Calderwood, L. (2016). Effects of Mode on Response Rates. Paper presented at the CLOSER Workshop: Mixing Modes and Measurement Methods in Longitudinal Studies, London.

Carpenter, H., and Burton, J. (2018). Adaptive Push-to-Web: Experiments in a Household Panel Study. Understanding Society Working Paper Series, No. 2018-05. https://www.understandingsociety.ac.uk/sites/default/files/downloads/working-papers/2018-05.pdf

Daikeler, J., Bošnjak, M., and Lozar Manfreda, K. (2019). Web Versus Other Survey Modes: An Updated and Extended Meta-Analysis Comparing Response Rates. *Journal of Survey Statistics and Methodology*, advance online access.

De Leeuw, E. D. (2018). Mixed-Mode: Past, Present, and Future. *Survey Research Methods*, 12(2), 75-89.

De Leeuw, E. D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, 21(5), 233-255.

Dillman, D., Smyth, J. & Christian, L.M. (2009). *Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method*, 3rd edition. Hoboken, New Jersey: Wiley.

Elliott, M. N., Zaslavsky, A. M., Goldstein, E., Lehrman, W., Hambarsoomians, K., Beckett, M. K., and Giordano, L. (2009). Effects of Survey Mode, Patient Mix, and Nonresponse on CAHPS Hospital Survey Scores. *Health Services Research*, 44(2), 501-518.

Fong, B., & Williams, J. (2011). British Crime Survey: Feasibility of Boosting Police Force Area (PFA) Sample Sizes Using Supplementary Recontact Surveys *Report for the Home Office*. London: TNS-BMRB.

Gaia, A. (2017). The Effect of Respondent Incentives on Panel Attrition in a Sequential Mixed-Mode Design. No. 2017-03, UK: Institute of Social and Economic Research.

Greene, J., Speizer, H., and Wiitala, W. (2008). Telephone and Web: Mixed-Mode Challenge. *Health Services Research*, 43(1), 230-248.

Heerwegh, D., and 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.

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.

Hochstim, J. R. (1967). A Critical Comparison of Three Strategies of Collecting Data from Households. *Journal of the American Statistical Association*, 62(319), 976–989.

Hope, S., Campanelli, P., Nicolaas, G., Lynn, P., and Jäckle, A. (2014). The Role of the Interviewer in Producing Mode Effects: Results from a Mixed Modes Experiment Comparing

Face-to-Face, Telephone and Web Administration. No. 2014-20, ISER Working Paper Series, available at https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2014-20.pdf.

Imbens, G. W. & Rubin, D. B. (2015). Instrumental Variables Analysis of Randomized Experiments with One-sided Noncompliance, Chapter 24 of Causal Inference for Statistics, Social, and Biomedical Sciences, an Introduction. Cambridge University Press. Chapter DOI: http://dx.doi.org/10.1017/CBO9781139025751.025

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.

Jäckle, A., Roberts, C., Lynn, P. (2010). Assessing the Effect of Data Collection Mode on Measurement. *International Statistical Review*, 78(1), 3–20.

Jäckle, A. (2016). Identifying and Predicting the Effects of Data Collection Mode on Measurement. Paper presented at the CLOSER Mixing Modes and Measurement Methods in Longitudinal Studies Workshop, London.

Jäckle, A., Gaia, A., and Benzeval, M. (2017). Mixing Modes and Measurement Methods in Longitudinal Studies. Report Submitted to the Cohort & Longitudinal Studies Enhancement Resources (CLOSER). UCL Institute of Education, London.

Kappelhof, J. (2015). The Impact of Face-to-Face vs Sequential Mixed-Mode Designs on the Possibility of Nonresponse Bias in Surveys Among Non-Western Minorities in the Netherlands. *Journal of Official Statistics*, 31(1), 1–31.

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.

Laaksonen, S., and Heiskanen, M. (2014). Comparison of Three Modes for a Crime Victimization Survey. *Journal of Survey Statistics and Methodology*, 2(4), 459-483.

Lozar Manfreda, K. & Vehovar, V. (2002). Mode Effect in Web Surveys. In the proceedings from The American Association for Public Opinion Research (AAPOR) 57th Annual Conference, 2002.

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., and Vehovar, V. (2008). Web Surveys Versus Other Survey Modes: A Meta-Analysis Comparing Response Rates. *International Journal of Market Research*, 50(1), 79-104.

McHorney, C. A., Kosinski, M., and Ware, J.E. Jr. (1994). Comparisons of the Costs and Quality of Norms for the SF-36 Health Survey Collected by Mail versus Telephone Interview: Results from a National Survey. *Medical Care*, 32(6), 551–567.

McMorris, B. J., Petrie, R. S., Catalano, R. F., Fleming, C. B., Haggerty, K. P., and Abbott, R. D. (2009). Use of Web and in-Person Survey Modes to Gather Data from Young Adults on Sex and Drug Use an Evaluation of Cost, Time, and Survey Error Based on a Randomized Mixed Mode Design. *Evaluation Review*, 33(2), 138–158.

Sakshaug, J. W., Cernat, A., and Raghunathan, T. E. (2019). Do Sequential Mixed-Mode Surveys Decrease Nonresponse Bias, Measurement Error Bias, and Total Bias? An Experimental Study. *Journal of Survey Statistics and Methodology*, online advance access.

Schonlau, M., Zapert, K., Simon, L., Sanstad, K., Marcus, S., Adams, J., Spranca, M., et al. (2003). A Comparison Between Responses From a Propensity-Weighted Web Survey and an Identical RDD Survey. *Social Science Computer Review*, 22(1), 128-138.

Scott, A., Jeon, S. H., Joyce, C. M., Humphreys, J. S., Kalb, G., Witt, J., and Leahy, A. (2011). A Randomised Trial and Economic Evaluation of the Effect of Response Mode on Response Rate, Response Bias, and item Non-Response in a Survey of Doctors. *BMC Medical Research Methodology*, 11(1), 126.

Siemiatycki, J. (1979). A Comparison of Mail, Telephone, and Home Interview Strategies for Household Health Surveys. *American Journal of Public Health*, 69(3), 238–245.

Wagner, J., Arrieta, J., Guyer, H., and Ofstedal, M. B. (2014). Does Sequence Matter in Multimode Surveys: Results from an Experiment. *Field Methods*, 26(2), 141–155.

Wood, M., & Kunz, S. (2014). CAWI in a Mixed Mode Longitudinal Design *Understanding Society Working Paper Series 2014-07*. Colchester: University of Essex.

Ye, C., Fulton, J., and Tourangeau, R. (2011). More Positive or More Extreme? A Meta-Analysis of Mode Differences in Response Choice. *Public Opinion Quarterly*, 75(2), 349-365.

Tables

Table 1: Definitions of key model variables

Variable in model	Definition	Note
W_i	Response by web	Treatment of principal interest
Y_i	Outcome of interest	Participation in the study sweep (1-0) Item responses (1-0) Item values (can be discrete or continuous)
Z_i	Random assignment to the mixed mode treatment group	This causally effects response by web, since only those allocated to the mixed mode group can respond by web

Table 2: Respondent profiles – all mixed mode vs all telephone-only (n = 10,586 cohort members randomised into experimental groups)

		Mean		Mean difference		
	Number non-missing	Mixed mode (Z=1) (n = 9110)	Telephone-only (Z=0) (n = 1476)	Estimate	95% CI	p-value
Whether participated at 50	10,586	0.88	0.88	0.00	-0.02, 0.02	0.93
Whether email provided to study	10,586	0.72	0.72	0.00	-0.02, 0.03	0.94
Computer use – age 50						
Computer at home	10,586	0.89	0.89	-0.01	-0.03, 0.01	0.37
Use home computer >2 per week	10,586	0.62	0.62	0.01	-0.02, 0.03	0.62
Personal access to internet	10,586	0.76	0.77	-0.01	-0.03, 0.02	0.54
Computer skills excellent or good	10,586	0.34	0.36	-0.02	-0.04, 0.01	0.23
Demographics – age 50						
Sex	10,586	0.49	0.49	0.00	-0.03, 0.03	0.93
White British	10,586	0.96	0.95	0.01	-0.01, 0.02	0.33
Homeowner	10,586	0.83	0.83	-0.01	-0.03, 0.01	0.47
In work	10,586	0.84	0.84	0.00	-0.03, 0.03	0.94
Professional/managerial occupation	-[A]	0.45	0.44	0.00	-0.03, 0.04	0.79
Weekly net pay	-[B]	405	402	3	-47, 53	0.91
Living with partner	10,586	0.79	0.78	0.01	-0.02, 0.03	0.52
Has degree	10,586	0.33	0.35	-0.02	-0.05, 0.01	0.18
Health and health behaviours – age 50						
Fair or poor health	10,586	0.19	0.18	0.02	-0.01, 0.04	0.19
Regular smoker	10,586	0.23	0.22	0.01	-0.01, 0.04	0.33
Problematic drinker	10,586	0.19	0.17	0.02	-0.01, 0.04	0.19
CASP-12 Quality of Life Score	10,586	25.9	26.1	-0.2	-0.5, 0.2	0.33
Cognitive function – age 50						
Animal naming test score	10,586	22.1	22.2	-0.1	-0.5, 0.3	0.52
Word list recall score	10,586	6.5	6.5	0.0	-0.1, 0.0	0.28
Delayed word list recall score	10,586	5.4	5.4	-0.1	-0.2, 0.1	0.29
Letter cancellation speed score	10,586	26.0	26.2	-0.2	-0.6, 0.2	0.34
Letter cancellation accuracy score (low score=greater accuracy)	10,586	4.5	4.6	-0.1	-0.4, 0.1	0.22

CI: Confidence interval.

[A] Numbers of cohort members in employment varied between 8890 and 8931 across imputed datasets.

[B] Numbers of cohort members in employment and with non-zero weekly net pay varied between 7616 and 7655 across imputed datasets.

Imputation model included experimental group assignment and all the variables in the table plus variables predictive of non-response (social class at birth,

region of residence at birth, Rutter behavioural score at age 7, social environment at age 7, region of residence at age 11, special educational needs at age 11, English ability at age 11 and general ability test score at age 11).
Joint likelihood ratio test of all parameters relating to the above variables (excluding weekly net pay and professional/managerial occupation, which are not defined on the whole sample so would reduce the analysis sample size) in a probit model with randomised group as the outcome gives p = 0.86.

Table 3: Survey response

	Experiment (n = 10,586)		Excluded from experiment (n = 967)			Total
	Mixed mode (Z = 1)	Tele-phone-only (Z = 0)	Emigrants (no UK telephone number)	UK Cases with no telephone number	Other	
Issued	9110	1476	367	572	28	11553
Inter-viewed	7547	1149	194	44	25	8959
Response Rate	82.8%	77.8%	52.9%	7.7%	89.3%	77.5%

Note: Data from the 178 'Dress Rehearsal' interviews (all by mixed mode, with 137 web and 41 telephone interviews) has been merged with the data collected during the main stage of data collection giving a total of 9,137 interviews in the Age 55 NCDS (Study Number 7669) deposited data at UKDA. A more detailed breakdown of survey response is provided in the NCDS 2013 Follow-Up Technical Report (www.cls.ioe.ac.uk/ncds9techreport).

Table 4: Compliance status – response by web or telephone

	Mixed mode (Z = 1) (n = 9110)			Telephone-only (Z = 0) (n = 1476)	
	Responded				
	Web (W = 1)	Telephone (W = 0)	Non-re-sponse	Responded by tele-phone (W = 0)	Non-re-sponse
N	5612	1935	1563	1149	327
% of all in group	61.6%	21.2%	17.2%	77.8%	22.2%
% of respond-ers	74.4%	25.6%	-	100%	-

Eighty five per cent of web-responders used either a PC or a laptop to complete the questionnaire. Another 14% used a tablet and less than one per cent used some other device (note response by Smartphone was not permitted). Currently, the typical ratio for UK online surveys of the general population is 80% PC/laptop and 20% tablet/other device so the device profile for the NCDS was slightly 'older' in terms of technology.

Table 5: Respondent profiles – mixed mode web vs mixed mode telephone (n = 7,547 mixed mode respondents)

		Mean		Mean difference		
	Number non-missing	Mixed mode web (Z = 1, W = 1) (n = 5612)	Mixed mode telephone (Z = 1, W = 0) (n = 1935)	Estimate	95% CI	p-value
Whether participated at 50	7,547	0.95	0.90	0.04	0.03, 0.06	<0.001
Whether email provided to study	7,547	0.92	0.45	0.47	0.44, 0.49	<0.001
Computer use – age 50						
Computer at home	7,547	0.95	0.79	0.16	0.14, 0.18	<0.001
Use home computer >2 per week	7,547	0.74	0.43	0.32	0.29, 0.34	<0.001
Personal access to internet	7,547	0.88	0.59	0.29	0.27, 0.31	<0.001
Computer skills excellent or good	7,547	0.43	0.20	0.24	0.21, 0.26	<0.001
Demographics – age 50						
Sex	7,547	0.49	0.49	-0.00	-0.03, 0.02	0.82
White British	7,547	0.96	0.95	0.01	0.00, 0.02	0.01
Homeowner	7,547	0.89	0.74	0.15	0.13, 0.17	<0.001
In work	7,547	0.88	0.79	0.09	0.07, 0.11	<0.001
Professional/managerial occupation	-[A]	0.52	0.32	0.21	0.18, 0.23	<0.001
Weekly net pay	-[B]	440	340	100	45, 155	<0.001
Living with partner	7,547	0.84	0.73	0.11	0.09, 0.14	<0.001
Has degree	7,547	0.41	0.23	0.18	0.15, 0.20	<0.001
Health and health behaviours – age 50						
Fair or poor health	7,547	0.14	0.27	-0.13	-0.15, -	<0.001
Regular smoker	7,547	0.18	0.31	-0.13	-0.16, -	<0.001
Problematic drinker	7,547	0.17	0.21	-0.04	-0.07, -	<0.001
CASP-12 Quality of Life Score	7,547	26.6	25.0	1.5	1.2, 1.9	<0.001
Cognitive function – age 50						
Animal naming test score	7,547	23.0	20.9	2.0	1.7, 2.4	<0.001
Word list recall score	7,547	6.7	6.2	0.5	0.4, 0.6	<0.001
Delayed word list recall score	7,547	5.6	5.0	0.6	0.5, 0.7	<0.001
Letter cancellation speed score	7,547	26.2	25.4	0.8	0.4, 1.2	<0.001
Letter cancellation accuracy score (low score=greater accu-	7,547	4.2	4.9	-0.7	-0.9, -0.5	<0.001

CI: Confidence interval.

[A] Numbers of cohort members in employment varied between 8890 and 8931 across imputed datasets.

[B] Numbers of cohort members in employment and with non-zero weekly net pay varied between 7616 and 7655 across imputed datasets.
Imputation model included experimental group assignment and response status, and all the variables in the table plus variables predictive of non-response (social class at birth, region of residence at birth, Rutter behavioural score at age 7, social environment at age 7, region of residence at age 11, special educational needs at age 11, English ability at age 11 and general ability test score at age 11).
Joint likelihood ratio test of all parameters relating to the above variables (excluding weekly net pay and professional/managerial occupation, which are not defined on the whole sample so would reduce the analysis sample size) in a probit model with mixed mode web vs mixed mode telephone as the outcome gives $p < 0.001$.

Table 6: Responders vs non-responders in the mixed mode and telephone-only groups (n = 10,586 cohort members randomised into experimental groups)

		Mean				Mean difference (1) - (3)		
		Mixed mode (Z = 1) (n = 9110)		Telephone-only (Z = 0) (n = 1476)				
	Number non-missing	(1) Respondents (W = 0, 1) (n = 7547)	(2) Non-respondents (W ≠ 0, 1) (n = 1563)	(3) Respondents (W = 0) (n = 1149)	(4) Non- respondents (W ≠ 0) (n = 327)	Estimate	95% CI	p-value
Whether participated at 50	10,586	0.94	0.62	0.95	0.63	-0.02	-0.03,	0.04
Whether email provided to study	10,586	0.80	0.36	0.79	0.46	0.00	-0.02,	0.84
Computer use – age 50								
Computer at home	10,586	0.91	0.78	0.92	0.79	-0.01	-0.03,	0.11
Use home computer >2 per week	10,586	0.66	0.43	0.66	0.46	0.00	-0.03,	0.79
Personal access to internet	10,586	0.80	0.55	0.82	0.59	-0.02	-0.03,	0.22
Computer skills excellent or good	10,586	0.37	0.17	0.40	0.20	-0.03	-0.06,	0.10
Demographics – age 50								
Sex	10,586	0.49	0.53	0.48	0.54	0.01	-0.02,	0.60
White British	10,586	0.96	0.95	0.96	0.93	0.00	-0.01,	0.69
Homeowner	10,586	0.85	0.70	0.86	0.73	-0.01	-0.04,	0.31
In work	10,586	0.86	0.76	0.86	0.77	-0.01	-0.03,	0.60
Professional/managerial occupation	-[A]	0.48	0.28	0.48	0.30	0.00	-0.03,	0.94
Weekly net pay	-[B]	417	340	413	359	4	-56, 64	0.90
Living with partner	10,586	0.81	0.69	0.80	0.71	0.01	-0.02,	0.61
Has degree	10,586	0.36	0.18	0.38	0.24	-0.02	-0.05,	0.20
Health and health behaviours – age 50								
Fair or poor health	10,586	0.17	0.28	0.16	0.23	0.01	-0.01,	0.29
Regular smoker	10,586	0.21	0.35	0.19	0.32	0.02	-0.01,	0.18
Problematic drinker	10,586	0.18	0.24	0.15	0.26	0.03	0.01,	0.02
CASP-12 Quality of Life Score	10,586	26.2	24.7	26.4	25.0	-0.2	-0.6, 0.1	0.22
Cognitive function – age 50								
Animal naming test score	10,586	22.4	20.5	22.5	21.3	-0.1	-0.5, 0.3	0.74
Word list recall score	10,586	6.6	6.1	6.6	6.3	0.0	-0.1, 0.1	0.46
Delayed word list recall score	10,586	5.5	4.9	5.5	5.2	0.0	-0.2, 0.1	0.55

Letter cancellation speed score	10,586	26.0	25.7	26.3	25.7	-0.3	-0.8, 0.2	0.28
Letter cancellation accuracy score (low score=greater	10,586	4.4	5.0	4.5	5.1	-0.1	-0.4, 0.1	0.32

CI: Confidence interval.

[A] Numbers of cohort members in employment varied between 8890 and 8931 across imputed datasets.

[B] Numbers of cohort members in employment and with non-zero weekly net pay varied between 7616 and 7655 across imputed datasets.

Imputation model included experimental group assignment and response status, and all the variables in the table plus variables predictive of non-response (social class at birth, region of residence at birth, Rutter behavioural score at age 7, social environment at age 7, region of residence at age 11, special educational needs at age 11, English ability at age 11 and general ability test score at age 11).

Joint likelihood ratio test of all parameters relating to the above variables (excluding weekly net pay and professional/managerial occupation, which are not defined on the whole sample so would reduce the analysis sample size) in a probit model with mixed mode responders vs. telephone-only responders as the outcome gives p = 0.41.

Table 7: Response to the survey – estimates of randomised group effects (n = 10,586 cohort members randomised into experimental groups)

	Response %			ITT analysis					
				Marginal effect of MM (no control)			Marginal effect of MM (controlling for pre-treatment characteristics[A])		
	Number non-missing	MM (Z = 1)	TO (Z = 0)	Estimate	95% CI	p-value	Estimate	95% CI	p-value
Probability of response to survey	10,586	82.8	77.9	5.0	2.7, 7.3	<0.001	5.2	3.2, 7.2	<0.001

CI: Confidence interval; ITT: Intention to treat.

[A] Controlling for: whether participated at 50, whether email provided to study, computer at home, use home computer >2 per week, personal access to internet, computer skills excellent or good, sex, white British, homeowner, in work, professional/managerial occupation, weekly net pay, living with partner, has degree, fair or poor health, regular smoker, problematic drinker, CASP-12 Quality of Life Score, animal naming test score, word list recall score, delayed word list recall score, letter cancellation speed score, letter cancellation accuracy score.

Imputation model included experimental group assignment and response status, plus the control variables listed above and variables predictive of non-response (social class at birth, region of residence at birth, Rutter behavioural score at age 7, social environment at age 7, region of residence at age 11, special educational needs at age 11, English ability at age 11 and general ability test score at age 11).

Table 8 MI: Item non-response – estimates of mode effects (n = 8696 responders)

	Item non-response (%)			ITT analysis						CACE analysis		
				Marginal effect of MM (no control)			Marginal effect of MM (controlling for pre-treatment characteristics[A])					
	Number non-missing	Mixed mode responders (Z = 1, W = 0, 1) (n = 7547)	Telephone-only responders (Z = 0, W = 0) (n = 1149)	Esti-mate	95% CI	p-value	Esti-mate	95% CI	p-value	Esti-mate	95% CI	p-value
Expected value of prop-erty	7469	12.9	7.9	5.0	3.1, 6.8	<0.001	4.9	3.1, 6.8	<0.001	6.4	3.6, 9.1	<0.001
Amount to pay off on property	3391	15.5	11.1	4.5	1.3, 7.7	0.01	4.3	1.1, 7.4	0.01	5.4	1.0, 9.8	0.02
Gross weekly income	5579	13.9	10.5	3.5	1.0, 5.9	0.01	3.4	1.0, 5.7	0.01	4.3	0.9, 7.7	0.01
Number of cigarettes a day usually smoked	1206	2.3	0.0	2.3	1.4, 3.2	0.06	-[B]	-[B]	-[B]	-[B]	-[B]	-[B]
Derived weight – kg	8696	7.2	5.7	1.5	0.0, 2.9	0.05	1.4	-0.1, 2.9	0.06	1.9	-0.2, 4.0	0.08
Whether voted in last general election	8600	0.8	0.6	0.2	-0.3, 0.7	0.52	0.1	-0.4, 0.6	0.66	0.2	-0.5, 0.9	0.56
Units of alcohol con-sumed in last 7 days	6595	1.7	1.6	0.1	-0.7, 1.0	0.74	0.0	-0.9, 0.9	0.97	0.0	-1.2, 1.2	0.98
Likelihood of working at the age of 60	8649	2.2	2.0	0.1	-0.7, 1.0	0.74	0.1	-0.8, 0.9	0.89	0.3	-1.5, 0.9	0.60
Frequency of alcohol consumption	8609	0.1	0.1	0.1	-0.1, 0.3	0.54	-[B]	-[B]	-[B]	-[B]	-[B]	-[B]
Likelihood of working at the age of 66	8645	2.2	2.4	-0.2	-1.1, 0.8	0.72	-0.2	-1.1, 0.7	0.68	-0.3	-1.5, 0.9	0.60
Party voted for in 2010 general election	6305	7.7	8.0	-0.4	-2.3, 1.6	0.72	-0.4	-2.4, 1.6	0.69	-0.5	-3.0, 2.0	0.69
Employer provided pen-sion Type A or Type B	3753	5.5	13.1	-7.6	-10.7, -4.6	<0.001	-8.0	-10.9, -5.0	<0.001	-9.7	-12.5, 6.9	<0.001

ITT: Intention to treat; CACE: Complier average causal effect; CI: Confidence interval.

[A] Controlling for: whether participated at 50, whether email provided to study, computer at home, use home computer >2 per week, personal access to inter-net, computer skills excellent or good, sex, white British, homeowner, in work, professional/managerial occupation, weekly net pay, living with partner, has degree, fair or poor health, regular smoker, problematic drinker, CASP-12 Quality of Life Score, animal naming test score, word list recall score, delayed word list recall score, letter cancellation speed score, letter cancellation accuracy score.

[B] Inestimable due to very low item non-response among one or both groups.

Imputation model included experimental group assignment and response status, and item non-response indicators for all the variables in the table plus the control variables listed above and variables predictive of non-response (social class at birth, region of residence at birth, Rutter behavioural score at age 7,

social environment at age 7, region of residence at age 11, special educational needs at age 11, English ability at age 11 and general ability test score at age 11).

Table 9 MI: Mean estimates – mixed mode vs telephone-only (n = 8696 responders)

	Mean			ITT analysis						CACE analysis		
				Marginal effect of MM (no control)			Marginal effect of MM (controlling for pre-treatment characteristics[A])					
	Number non-missing	Mixed mode respond-ers (Z = 1, W = 0, 1) (n = 7547)	Telephone-only re-sponders (Z = 0, W = 0) (n = 1149)	Esti-mate	95% CI	p-value	Esti-mate	95% CI	p-value	CACE	95% CI	p-value
Socio-economic char-acteristics												
Whether working at 55	8574	0.81	0.82	-0.01	-0.04, 0.01	0.30	-0.01	-0.03, 0.01	0.45	-0.01	-0.04, 0.02	0.41
Whether profes-sional/managerial occu-pation at 55	8696	0.35	0.35	0.00	-0.03, 0.03	0.91	0.01	-0.01, 0.04	0.34	0.02	-0.02, 0.05	0.38
Hours worked per week in main job	5503	37.1	37.7	-0.6	-1.5, 0.4	0.27	-0.6	-1.5, -0.2	0.15	-0.8	-2.0, 0.3	0.17
Gross weekly pay (if has a job) – capped £4000	4828	623	593	30	-18, 77	0.23	54	-11, 119	0.11	67	-11, 146	0.09
Net weekly pay (if has a job) – capped £2500	4775	425	411	14	-12, 40	0.29	23	-16, 61	0.24	25	-19, 69	0.26
Age 55 - self-rated fi-nancial situation - re-verse coded	8586	3.83	4.00	-0.17	-0.23, -0.10	<0.001	-0.15	-0.21, -0.09	<0.001	-0.17	-0.25, -0.10	<0.001
Any qualifications re-ported	8604	0.15	0.14	0.01	-0.01, 0.03	0.29	0.01	-0.01, 0.04	0.22	0.02	-0.01, 0.05	0.24
Number of qualifications reported	8696	0.17	0.15	0.02	0.00, 0.05	0.08	0.03	0.00, 0.05	0.06	0.03	0.00, 0.07	0.08
Likelihood of working at the age of 60	8464	68.7	66.7	2.0	-0.3, 4.4	0.09	2.1	0.0, 4.3	0.06	2.9	-0.1, 5.8	0.06
Likelihood of working at the age of 66	8454	39.0	35.4	3.6	1.2, 6.0	0.004	3.6	1.2, 6.0	0.003	4.5	1.5, 7.6	0.004
Expected value of prop-erty	6558	373,356	404,229	-30,874	-71,597, 9850	0.14	-33,518	-73,566, 6530	0.10	-43,472	-91,953, 5010	0.08
Amount yet to pay off on property	2884	120,835	106,193	14,643	-23,888, 53,174	0.46	14,595	-23,957, 53,147	0.46	18,342	-29,917, 66,601	0.46
Number of relationships reported	8692	1.07	1.07	0.00	-0.01, 0.02	0.68	0.00	-0.02, 0.02	0.98	0.00	-0.02, 0.02	0.78

Number of addresses reported	1566	1.39	1.35	0.04	-0.07, 0.14	0.50	0.03	-0.07, 0.14	0.55	0.03	-0.11, 0.16	0.70
Number of economic activities	2728	1.52	1.51	0.01	-0.10, 0.12	0.90	0.00	-0.11, 0.11	0.99	0.00	-0.13, 0.14	0.95
Health												
Self-rated general health (1 to 5) reverse coded	8612	3.33	3.43	-0.10	-0.17, -0.04	0.003	-0.07	-0.13, -0.01	0.02	-0.08	-0.15, 0.00	0.04
Whether classified as disabled	8579	0.20	0.17	0.03	0.01, 0.06	0.004	0.03	0.01, 0.05	0.01	0.04	0.01, 0.07	0.02
Whether has a long-standing illness	8590	0.33	0.30	0.03	0.00, 0.06	0.03	0.03	0.00, 0.05	0.05	0.04	0.00, 0.07	0.06
Number of health problems reported	8696	1.24	1.11	0.13	0.05, 0.21	0.001	0.11	0.04, 0.18	0.003	0.14	0.05, 0.24	0.003
(Derived) Weight in kilograms	8085	79.4	79.8	-0.4	-1.5, 0.8	0.51	-0.6	-1.6, 0.4	0.22	-0.9	-2.3, 0.5	0.21
Specific health problems												
Asthma or wheezy bronchitis	8593	0.12	0.11	0.01	-0.01, 0.03	0.23	0.01	-0.01, 0.03	0.41	0.01	-0.01, 0.04	0.36
Diabetes	8598	0.07	0.06	0.00	-0.01, 0.02	0.61	0.00	-0.01, 0.02	0.81	0.00	-0.02, 0.02	0.71
Backache, sciatica, disc prolapse	8596	0.26	0.23	0.03	0.00, 0.05	0.04	0.03	0.00, 0.05	0.05	0.03	0.00, 0.07	0.07
Cancer or leukaemia	8599	0.04	0.04	0.00	-0.01, 0.01	0.90	0.00	-0.01, 0.01	0.80	0.00	-0.02, 0.01	0.84
Problems with hearing	8598	0.11	0.09	0.02	0.00, 0.04	0.03	0.02	0.00, 0.04	0.04	0.02	0.00, 0.05	0.06
High blood pressure	8595	0.22	0.22	0.00	-0.02, 0.03	0.78	0.00	-0.03, 0.02	0.93	0.00	-0.04, 0.03	0.93
Heart problems	8586	0.06	0.04	0.01	0.00, 0.03	0.04	0.01	0.00, 0.03	0.06	0.01	0.00, 0.03	0.12
Depression, emotional and psychiatric	8590	0.16	0.13	0.02	0.00, 0.05	0.03	0.02	0.00, 0.04	0.04	0.02	0.00, 0.05	0.09
Health behaviours												
Number of units of alcohol consumed in last 7 days	6484	10.6	8.1	2.5	1.6, 3.4	<0.001	1.9	1.2, 2.7	<0.001	2.4	1.5, 3.3	<0.001
Whether drinks most days	8597	0.18	0.17	0.01	-0.01, 0.03	0.37	0.00	-0.02, 0.03	0.81	0.00	-0.03, 0.04	0.77
Whether regular smoker	8597	0.14	0.13	0.01	-0.01, 0.03	0.46	0.00	-0.02, 0.01	0.62	0.00	-0.03, 0.02	0.74

Well-being - all recoded so higher scores = better well-being												
My age prevents me from doing the things I would like to do	8580	3.13	3.22	-0.09	-0.15, -0.04	0.001	-0.07	-0.13, -0.02	0.01	-0.10	-0.17, -0.03	0.004
I feel what happens to me is out of my control	8566	3.07	3.13	-0.06	-0.12, -0.01	0.03	-0.05	-0.10, 0.00	0.07	-0.06	-0.13, 0.01	0.09
I feel left out of things	8572	3.28	3.42	-0.15	-0.20, -0.09	<0.001	-0.14	-0.19, -0.09	<0.001	-0.18	-0.24, -0.12	<0.001
I feel full of energy these days	8584	2.91	3.00	-0.10	-0.15, -0.04	0.001	-0.08	-0.13, -0.03	0.002	-0.10	-0.17, -0.04	0.002
I feel that life is full of opportunities	8573	3.03	3.15	-0.12	-0.17, -0.06	<0.001	-0.10	-0.15, -0.05	<0.001	-0.14	-0.20, -0.08	<0.001
I feel that the future looks good for me	8538	3.20	3.36	-0.16	-0.21, -0.11	<0.001	-0.14	-0.18, -0.09	<0.001	-0.19	-0.25, -0.13	<0.001
Leisure												
Play sport or go walking or swimming at least once a week	8585	0.63	0.69	-0.07	-0.09, -0.04	<0.001	-0.06	-0.09, -0.03	<0.001	-0.08	-0.12, -0.04	<0.001
Have a meal in a pub or restaurant at least once a week	8589	0.19	0.27	-0.07	-0.10, -0.05	<0.001	-0.07	-0.10, -0.05	<0.001	-0.10	-0.13, -0.07	<0.001
Voting												
Whether voted conservative 2010	5818	0.37	0.40	-0.02	-0.06, 0.01	0.20	-0.02	-0.06, 0.01	0.25	-0.03	-0.07, 0.02	0.26
Whether voted labour 2010	5818	0.32	0.35	-0.03	-0.06, 0.01	0.12	-0.03	-0.07, 0.00	0.09	-0.04	-0.08, 0.01	0.09
Whether voted liberal democrat 2010	5818	0.19	0.15	0.04	0.01, 0.07	0.01	0.04	0.01, 0.06	0.01	0.05	0.01, 0.08	0.01

ITT: Intention to treat; CACE: Complier average causal effect; CI: Confidence interval.

[A]Controlling for: whether participated at 50, whether email provided to study, computer at home, use home computer >2 per week, personal access to internet, computer skills excellent or good, sex, white British, homeowner, in work, professional/managerial occupation, weekly net pay, living with partner, has degree, fair or poor health, regular smoker, problematic drinker, CASP-12 Quality of Life Score, animal naming test score, word list recall score, delayed word list recall score, letter cancellation speed score, letter cancellation accuracy score.

Imputation model included experimental group assignment and response status, and all the variables in the table plus the control variables listed above and variables predictive of non-response (social class at birth, region of residence at birth, Rutter behavioural score at age 7, social environment at age 7, region of residence at age 11, special educational needs at age 11, English ability at age 11 and general ability test score at age 11).

Figures

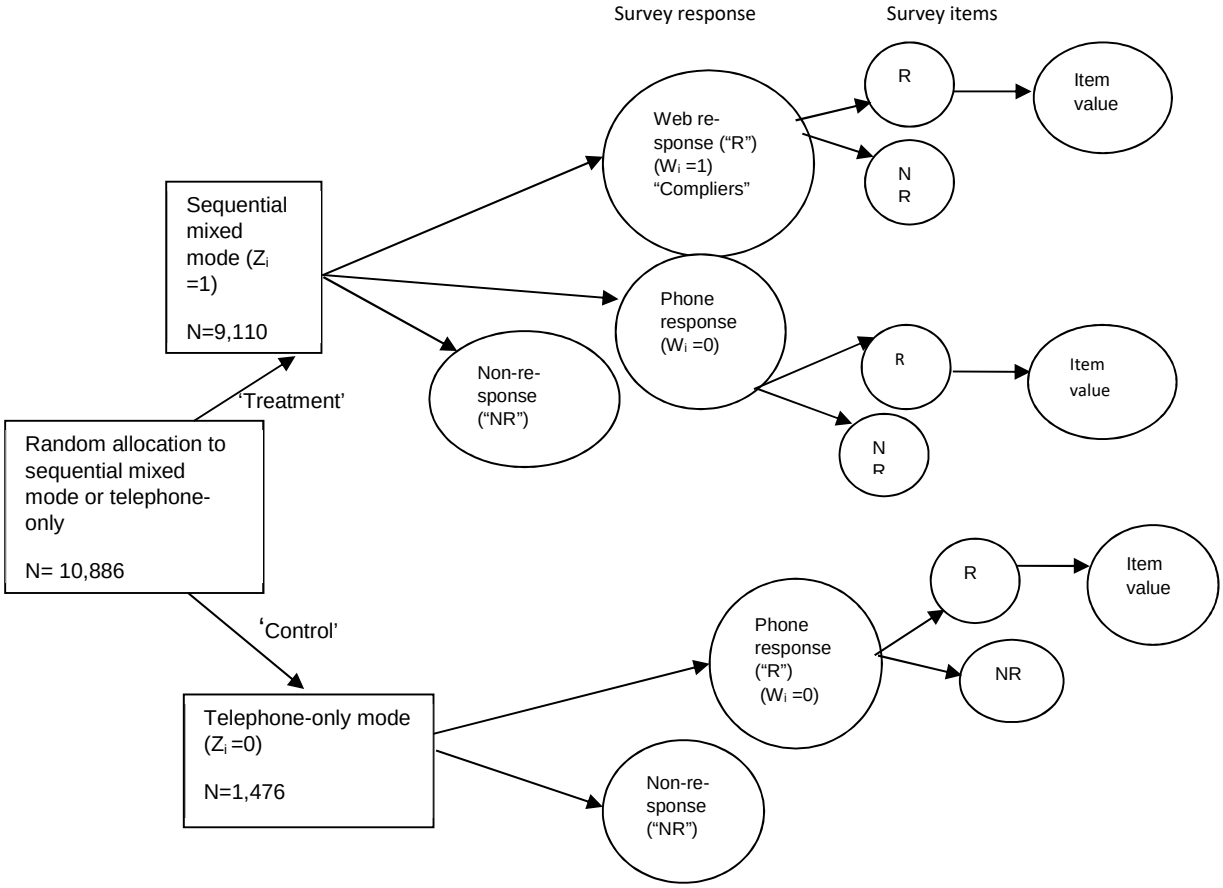

Figure 1: Structure of the mixed mode experiment