Understanding Society

Working Paper Series

No. 2015-03

May 2015

Understanding Society Innovation Panel Wave 7:

Results from Methodological Experiments

Tarek Al Baghal (ed.)

Contributors: Annalies G. Blom[1], Jonathan Burton[4], Cara Booker[4], Alexandru Cernat[4], Malcolm Fairbrother[5], Annette Jäckle[4], Olena Kaminska[4], Florian Keusch[2], Jon A. Krosnick[3], Peter Lynn[4], Daniel Oberski[6], Stephen Pudney[4], Emanuela Sala[7], Sebastian Schnettler[9], Henning Silber[8], Tobias H. Stark[10], S.C. Noah Uhrig[4], Ting Yan[2]

[1] University of Mannheim, [2] University of Michigan, [3] Stanford University, [4] Institute of Social and Economic Research, University of Essex, [5] University of Bristol, [6] Tilburg University, [7]University of Milan-Bicocca, [8] University of Goettingen, [9] University of Konstanz, [10] Utrecht University

Understanding Society Innovation Panel Wave 6: Results from Methodological Experiments

Tarek Al Baghal (ed.)

Contributors: Annalies G. Blom[1], Jonathan Burton[4], Cara Booker[4], Alexandru Cernat[4], Malcolm Fairbrother[5], Annette Jäckle[4], Olena Kaminska[4], Florian Keusch[2], Jon A. Krosnick[3], Peter Lynn[4], Daniel Oberski[6], Stephen Pudney[4], Emanuela Sala[7], Sebastian Schnettler[9], Henning Silber[8], Tobias H. Stark[4], S.C. Noah Uhrig[4], Ting Yan[2]

[1] University of Mannheim, [2] University of Michigan, [3] Stanford University, [4] Institute of Social and Economic Research, University of Essex, [5] University of Bristol, [6] Tilburg University, [7]University of Milan-Bicocca, [8] University of Goettingen, [9] University of Konstanz, [10] Utrecht University

Non-technical summary

The *Understanding Society* survey includes what is known as an 'Innovation Panel' sample (IP). This sample of originally 1500 households is used to test different methods for conducting longitudinal surveys in order to produce the highest quality data. The results from the Innovation Panel provide evidence about the best way to conduct a longitudinal survey which is of relevance for all survey practitioners as well as influencing decisions made about how to conduct *Understanding Society*. This paper reports the experiments with the mixed- mode design and early results of the methodological tests carried out at wave 7 of the Innovation Panel in the summer and fall of 2014.

To bolster the number of responding households, a refreshment sample was included at IP7. IP7 was also the third wave employing a mixed-mode design including an internet survey, and the fourth wave of the Innovation Panel to employ a mixed-mode design generally. IP2 had experimented with telephone interviewing in addition to face-to-face personal interviewing. Like IP5 and IP6, IP7 uses a design in which households were allocated to a sequential mixed-mode design. This allocation only includes households in the sample prior to IP7, and the IP7 refreshment sample to part only in face-to-face interviews. The adults in the mixed-mode design were first approached by letter and email where possible and asked to complete their interview on-line. Those who did not respond on-line were then followed up by face-to-face interviewers. The remaining households from older samples were issued directly to face-to-face interviewers.

The methodological tests included an experiment testing the effects of changing the amount of incentives offered to respondents in advance of fieldwork on response rates, the impact of making the incentive conditional or unconditional, the multiple intra-wave contacts. Further experiments examine the measurement of attitudes towards an environmental tax, the effect of question wording on response outcomes, the use of reminders in the question, the impact of how scales are presented, the measurement of finger length as an indicator of health outcomes, and the impact of motivational statements.

Understanding Society Innovation Panel Wave 7: Results from Methodological Experiments

Tarek Al Baghal (ed.)

Contributors: Annalies G. Blom[1], Jonathan Burton[4], Cara Booker[4], Alexandru Cernat[4], Malcolm Fairbrother[5], Annette Jäckle[4], Olena Kaminska[4], Florian Keusch[2], Jon A. Krosnick[3], Peter Lynn[4], Daniel Oberski[6], Stephen Pudney[4], Emanuela Sala[7], Sebastian Schnettler[9], Henning Silber[8], Tobias H. Stark[4], S.C. Noah Uhrig[4], Ting Yan[2]

[1] University of Mannheim, [2] University of Michigan, [3] Stanford University, [4] Institute of Social and Economic Research, University of Essex, [5] University of Bristol, [6] Tilburg University, [7]University of Milan-Bicocca, [8] University of Goettingen, [9] University of Konstanz, [10] Utrecht University

Abstract

This paper presents some preliminary findings from Wave 7 of the Innovation Panel (IP7) of *Understanding Society*: The UK Household Longitudinal Study. *Understanding Society* is a major panel survey in the UK. In June 2014, the seventh wave of the Innovation Panel went into the field. IP7 includes a new refreshment sample and used a mixed-mode design, using on-line interviews and face-to-face interviews. This paper describes the design of IP7, the experiments carried and the preliminary findings from early analysis of the data.

Key words: longitudinal, survey methodology, experimental design, respondent incentives, questionnaire design.

JEL classification: C80, C81, C83

Contact: Tarek Al Baghal (talbag@essex.ac.uk) Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK.

Table of Contents

1. Introduction

This paper presents early findings from the seventh wave of the Innovation Panel (IP7) of *Understanding Society*: The UK Household Longitudinal Study (UKHLS). *Understanding Society* is a major panel survey for the UK. The first five waves of data collection on the main sample have been completed, and sixth and seventh waves are currently in the field. The data from the first four waves of the main samples are available from the UK Data Archive, and the fifth will be available towards the end of 2015. Data from a nurse visit to collect bio-markers from the general population sample and the British Household Panel Survey (BHPS) are also available. Data for the first six waves of the Innovation Panel are available from the UK Data Service[1].

One of the features of *Understanding Society*, alongside the large sample size (40,000 households at Wave 1), the ethnic minority boost sample and the collection of bio-markers, is the desire to be innovative. This has been a key element of the design of *Understanding Society* since it was first proposed. Part of this drive for innovation is embodied within the Innovation Panel (IP). This panel of almost 1500 households was first interviewed in the early months of 2008. The design in terms of the questionnaire content and sample following rules are modelled on *Understanding Society*. The IP is used for methodological testing and experimentation that would not be feasible on the main sample. The IP is used to test different fieldwork designs, new questions and new ways of asking existing questions.

The second wave of the Innovation Panel (IP2) was carried out in April-June 2009, the third wave (IP3) in April-June 2010 and the fourth wave in March-July 2011. The fourth wave of the Innovation Panel (IP4) included a refreshment sample of 465 responding households. In March 2012, IP5 was fielded, with part of the samples conducting the survey via the internet, while others continued in an interviewer-administered survey. Fieldwork for IP6 started in March 2013, repeating the design where some were first asked to complete the survey via the web option while others were approached by an interviewer only. The IP6 also included a

[1] http://discover.ukdataservice.ac.uk/series/?sn=2000053

mop-up follow-up phase with anyone not responding with contacts attempted by CATI or CAWI at the end of the fieldwork. Working Papers which cover the experimentation carried out in all six innovation panels are available from the *Understanding Society* website.[2] The data from the first six waves of the innovation panel are held at the UK Data Service. This paper describes the design of IP7, the experiments carried and some preliminary findings from early analysis of the data. Section 2 outlines the main design features of *Understanding Society*. Section 3 describes the design and conduct of IP7. Section 4 then reports on the experiments carried at IP7.

2. *Understanding Society*: the UKHLS

Understanding Society is an initiative of the Economic and Social Research Council (ESRC) and is one of the major investments in social science in the UK. The study is managed by the Scientific Leadership Team (SLT), based at ISER at the University of Essex and including members from the University of Warwick and the London School of Economics. The fieldwork and delivery of the survey data for the first five waves of the main samples were undertaken by NatCen Social Research (NatCen). Waves 6 through 8 are being carried out by TNS-BMRB. *Understanding Society* aims to be the largest survey of its kind in the world. The sample covers the whole of the UK, including Northern Ireland and the Highlands and Islands of Scotland. *Understanding Society* provides high quality, longitudinal survey data for academic and policy research across different disciplines. The use of geo-coded linked data enables greater research on neighbourhood and area effects, whilst the introduction of bio-markers and physical measurements (Waves 2 and 3) opens up the survey to health analysts.

The design of the main-stage of *Understanding Society* is similar to that of the British

[2] https://www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2008-03
https://www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2010-04
https://www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2011-05
https://www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2012-06
https://www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2013-06
https://www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2014-04

Household Panel Survey (BHPS) and other national panels around the world. In the first wave of data collection, a sample of addresses was issued. Up to three dwelling units at each address were randomly selected, and then up to three households within each dwelling unit were randomly selected. Sample households were then contacted by NatCen interviewers and the membership of the household enumerated. Those aged 16 or over were eligible for a full adult interview, whilst those aged 10-15 were eligible for a youth self-completion. The adult interviews were conducted using computer-assisted personal interviewing (CAPI) using lap-tops running the questionnaire in Blaise software. Adults who participated in *Understanding Society* were also asked to complete a self-completion questionnaire, in which questions thought to be more sensitive were placed. The adult self-completions at Waves 1 and 2, and the youth self-completions, were paper questionnaires. From Wave 3 onwards the adult self-completion instrument was integrated into the interviewing instrument and the respondent used the interviewer's lap-top to complete that portion of the questionnaire themselves (Computer-Assisted Self-Interviewing, CASI).

In between each wave of data collection, sample members are sent short reports of early findings from the survey, and a change-of-address card, to allow them to inform ISER of any change in their address and contact details. Before each sample month is issued to field for a new wave, each adult is sent a letter which informs them about the new wave of a survey, includes a token of appreciation in the form of a gift voucher and also includes a change-of-address card. Interviewers then attempt to contact households and enumerate them, getting information of any new entrants into the household and the location of anyone who has moved from the household. New entrants are eligible for inclusion in the household. Those who move, within the UK, are traced and interviewed at their new address. Those people living with the sample member are also temporarily eligible for interview. More information about the sampling design of *Understanding Society* is available in Lynn (2009).[3] From Wave 2, the BHPS sample has been incorporated into the *Understanding Society* sample. The BHPS sample is interviewed in the first half of each wave.

[3]https://www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2009-01.pdf

3. Innovation Panel Wave 7: Design

A new refreshment sample was issued at IP7 to be included with the original sample from IP1 and the IP4 refreshment sample. IP7 employed a mixed-mode design, which started in IP5 and also employed in IP6. Starting at IP5, the modes which were mixed were on-line (CAWI) and face-to-face (CAPI) interviewing. In IP5, a random selection of two-thirds of households was allocated to the mixed-mode design ("WEB") with the remaining third of households allocated directly to face-to-face interviewers ("F2F"). This sample allocation was maintained at IP6 and IP7. The IP7 refreshment sample units were all allocated to the F2F design. TNS BMRB conducted fieldwork at IP7, after the first six waves were conducted by NatCen.

The sample was divided into two tranches. For one tranche (Tranche 1), the fieldwork for the WEB group started three weeks earlier than the F2F fieldwork. For the other tranche (Tranche 2), fieldwork for the WEB group started five weeks earlier than the F2F fieldwork. Initially, advance letters were sent to adults in the WEB group which included a URL and a unique log-in code. Adults in the WEB group for whom we had an email address were also sent an email which included a link which could be clicked through to the web-site. There were two email reminders for adults with an email address who had not yet completed their interview on-line, sent three days apart. A reminder letter was then sent to all adults in the WEB group who had not completed their interview. This letter was sent just under two weeks after the initial advance letter.

At the end of three or five weeks, all adults who had not completed their interview were allocated to face-to-face interviewers, but could still enter the web survey instead if they desired within the next four weeks of fieldwork. Adults who had started their interview on-line, but not reached the 'partial interview' marker, were issued to face-to-face interviewers. The interviewers were able to re-start the interview at the place at which the respondent had stopped. After these seven (Tranche 1) or nine weeks (Tranche 2), fieldwork for members in the WEB sample group stopped until the mop-up phase, and the F2F samples were issued to interviewers. CAPI fieldwork for the F2F sample in Tranche 1 lasted 16 weeks, and lasted 14 weeks for the F2F sample in Tranche 2. After these periods, the mop-up phase started.

The WEB-only period ran from 21st May to 12th June for Tranche 1 and 21st May to 24 the

June for Tranche 2. The face-to-face fieldwork for the WEB sample started 13th June and ran until 9th July for Tranche 1 and from 25th June to 24th July for Tranche 2. The face-to-face fieldwork for the F2F sample ran from 10th July to 19th October for Tranche 1 and 25th July to 19th October for Tranche 2 .The mop-up follow-up phase with those not responding in both the WEB and F2F versions, conducted through CATI with CAWI available was from 20th October to 2nd November for both tranches.

Prior to the survey going into the field there were eight half-day briefings for the interviewers. The briefings were conducted by TNS-BMRB researchers, with staff from ISER contributing to provide information about the study and to talk in more detail about the experiments. The locations of the briefings gave a wide geographic spread: London (three briefings), Warwick, Newcastle, Bristol, Manchester and Edinburgh. The briefings took place between 23rd May and 19th June 2014, with a total of 101 interviewers attending the briefings. A debrief also took place in September with a selection of interviewers from different areas. All interviewers working on the survey were provided with feedback forms and were asked to fill and return them to the TNS BMRB research team at the end of fieldwork.

a. Call for experiments

IP7 was the fifth time the Innovation Panel was open for researchers outside the scientific team of *Understanding Society* to propose experiments. A public call for proposals was made on 15th March 2013 with a deadline of 15th May. Eighteen proposals were received with eight being accepted, plus two carried over from previous waves (one from IP5, one from IP6). One additional module, measuring finger length, was included at IP7 for those adults not asked at IP6, including the new refreshment sample and 10-15 year olds. There were for a total of eleven experiments included in IP7, with many respondents taking part in only ten. The eighteen new submissions came from within ISER (eight), ISER in collaboration with other researchers (two) and from outside ISER completely (eight). The eighteen proposals were reviewed by a panel which included two ISER-based members of the *Understanding Society* scientific leadership team, and two members of the Methodology Advisory Committee to *Understanding Society* who were external to ISER. In addition to those experiments which were accepted through the public call, there were a number of core experiments which the Understanding Society senior leadership team wanted to run. These

core experiments included the mixed-mode design and the main incentives experiment.

In addition to these experiments, two Associated Studies were included in IP7, one of which was repeated from IP6. The repeated study is on time and risk preferences, aiming to combine survey data from IP6 and IP7 with experimental data on risk preferences (the attitude for taking a gamble) and time preferences (the degree to which today is valued more highly than tomorrow). A random selection of IP6 respondents was made, such that each household had only one individual selected to participate and 644 respondents answered these questions. At IP7, these same respondents were asked the same set of questions. One-tenth of these respondents were given a payment upon completion of the questions. Those selected to receive a payment were given an amount based on one of the 91 questions that they answered. Some of the questions involved a lottery, and a random mechanism was used to select which outcome of the lottery the respondent was paid.

The other Associated Study is a time use diary with respondents asked to detail their activities for a 24-hour period. All adults were asked to complete two diaries: one covering a week day and the other a weekend day. All household members were asked to complete a diary for the same days of the week. Households were randomly allocated to a day of the week and a weekend day. Where respondents were unable to complete on their allocated date, they were instructed to complete their diary on the same day in a subsequent week. All adults received an unconditional £5 incentive; this was given at the time the interviewers' handed over both diaries. Where a mobile number and/or email address was provided, respondents were sent text messages and/or emails, reminding them to fill in and return their time diaries. A telephone reminder stage was also included in an attempt to boost the diary completion rate. All respondents who had not returned their diary within two weeks of the completion date were included in the telephone reminder. 1,190 respondents completed and returned both diaries, with a further 73 completing the diary for one of their designated days. The overall rate of completion of the time diary among those asked was 55.1%.

b. Sample

There were three sample issued at IP7: the original sample; the refreshment sample from IP4; and a new refreshment sample issued at IP7. The original sample and IP4 refreshment sample at IP7 were comprised of those households who had responded at IP6, plus some households

which had not responded at IP6. Households which had adamantly refused or were deemed to be mentally or physically incapable of giving an interview were withdrawn from the sample. There were 965 original and 456 IP4 refreshment sample households issued at IP7.

After six waves of experimentation for the original sample and three waves for the IP4 refreshment sample, the sample size of the Innovation Panel had decreased to just over 1,000 households. The IP7 refreshment sample was designed to bring the productive sample up to the original size of 1,500 households. There were 1913 sample households issued for the IP7 refreshment sample. Like the other samples, the IP7 refreshment sample was selected from the Postcode Address File (PAF) from the 120 sectors used in the original sample.

As discussed above, around two-thirds of the original and IP4 refreshment samples were allocated to the mixed-mode design in IP5, which was maintained in IP6 and IP7. Sample members would be approached by letter and email (where possible) to complete their interview on-line. This experimental allocation did not include the IP7 refreshment sample, which were all allocated a face-to-face only design. The table below shows the allocation to mode design by sample type for those included in the issued original and IP4 refreshment samples in IP7.

Table 1: Allocation to mode design by sample type

	Original Sample	IP4 Refreshment Sample	Total
CAPI only	326 33.8%	166 36.4%	492 34.6%
Mixed-mode (CAWI+CAPI)	639 66.2%	290 63.6%	929 65.4%
Total	965	456	1,421

c. Questionnaire design

The questionnaire at IP7 followed the standard format used in the previous Innovation Panels as well as the main-stage of *Understanding Society*. The questionnaires used at IP7 are available from the *Understanding Society* website.[4] The interview included the following

[4] https://www.understandingsociety.ac.uk/documentation/innovation-panel/questionnaires

sections with the corresponding target times for each:

- Household roster and household questionnaire: 15 minutes per household
- Individual questionnaire: average 31 minutes for each person aged 16 or over
- Adult self-completion: around 9 minutes, paper questionnaire or computer self-administered interview (CASI)
- Youth self-completion: 10 minutes for each child aged 10-15 years
- Proxy questionnaire: 10 minutes for adults ages 16 or over who are not able to be interviewed.
- Time/Risk Preferences: 10 minutes for adults ages 16 or over who were selected for this study.

However, it was found during fielding that the survey was significantly longer than expected, and the estimated times were incorrect. The median survey time for the individual questionnaire using the original script was 63 minutes, significantly longer than the 50 minutes initially assumed. The survey was then adjusted to reduce completion time, with the Food Safety, Cognitive Ability, Fertility History, and Partnership History sections dropped from the individual questionnaire after the first six weeks of the field period. The removal of these sections brought the survey length back to the originally assumed length. The longer version of the survey, with these sections included, ran from 13th June to 22nd July. The shorter version, dropping these sections, ran from 23rd July to the end of the field period (19th October).

Some parts of the IP7 interview were recorded, using the lap-tops. This was done with permission of the respondent. The recordings were primarily of the experimental content of the questionnaire, and were taken to enable researchers to investigate the processes by which respondents came up with their answer. Almost 80% of those who participated agreed to the sound recording, which was higher in part waves; at IP4 recording consent was 68%. There was little difference in the levels of consent to recording between the original sample (79.3%), IP4 refreshment sample (82.6%), and IP7 refreshment sample (79.4%).

There were some changes made to the questionnaire to enable participants to complete it on-line at IP5 when the web design was first introduced, and can be described more in-depth in the working paper containing results from the experiments in IP5.[5] Briefly, the changes made to the questionnaire are as follows. Questions were reworded as needed to include interviewer instructions that may clarify the definition of the question. Text was altered to be more participant-focused rather than interviewer-focused. The first person in the household to log in to the web survey would be asked to complete the household enumeration. A question about who was responsible for paying household bills was included; the person or people indicated as responsible were routed first to the household questionnaire and then to the individual questionnaire.

If a participant had started to answer their questionnaire and left the computer for 10 minutes, they were automatically logged out. The participant was able to log back in using the same process as they had originally logged in, and they would be taken to the place that they had left the interview. This also applies to those who had closed down the browser mid-interview. A 'partial interview' marker was put into place about two-thirds of the way through the interview, after the benefits section. If a participant reached this stage, the interview was considered to be a 'partial interview'. They could log back in and complete if they wanted, but otherwise they were not contacted by an interviewer. If the participant had not reached this marker before closing down the browser, they were sent an email overnight which thanked them for their work so far and encouraged them to complete the survey, giving them the URL to click through to the survey. Again, they would start at the point where they had left off. In addition, those who had started but not reached the partial interview marker were, after the initial two weeks, issued to face-to-face interviewers who would be able to finish the survey with them, from where they had left off.

d. Response rates

This section sets out the response rates for IP7 as a whole. Section 4j describes the effect of incentives on response rates. Table 2 sets out the response rates for eligible households for the original, IP4 refreshment, and IP7 refreshment samples. In this and all following tables,

[5]https://www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2013-06

cells present both the percentage and the number of cases this percentage represents, while the bottom row presents total number of cases.

Table 2. Household response at IP7

	Original Sample	IP4 Refreshment Sample	IP7 Refreshment Sample	Total
Responding	78.8% 692	77.8% 325	33.6% 488	54.7% 1505
Non-contact	9.0% 79	9.8% 41	16.1% 234	12.9% 354
Refusals	10.5% 92	11.2% 47	45.1% 655	28.9% 794
Other non-responding	1.7% 15	1.2% 5	5.3% 77	3.5% 97
Total	878	418	1454	2750

Response for the IP7 refreshment sample was disappointing. We believe that there were a number of potential factors which may have contributed to a lower than anticipated response for this sample. This was the first Innovation Panel wave conducted by this fieldwork agency, and so was new for all the interviewers. Moreover, IP7 was a particularly complex survey, which included more experiments than previous IPs. In addition, there were a couple of Associated Studies which added to the complexity of the questionnaire, and to the time taken to complete the survey. Furthermore, as noted in section 3c, the questionnaire took longer to administer than estimated. This additional length, conveyed by the interviewer to the sample member when asked, may have affected response. The complexity would not have directly affected response at the household level, but may have contributed to within-household non-response, and is likely to have affected the motivation of interviewers.

Table 3 separates out the response rate for households that had responded at IP6 and those that had not. These are only pertinent to the original and IP4 refreshment samples, given that this wave was the first for those in the IP7 refreshment sample.

Table 3. Household response at IP7 by IP6 outcome

	Original Sample		IP4 Refreshment Sample	
	IP6 Responding	IP6 Non-Responding	IP6 Responding	IP6 Non-Responding
Responding	83.9% 662	33.7% 30	82.0% 287	55.6% 38
Non-contact	6.2% 49	33.7% 30	7.1% 25	23.5% 16
Refusals	8.4% 66	29.2% 26	9.7% 34	19.1% 13
Other non-responding	1.5% 12	3.4% 3	1.1% 4	1.5% 1
Total	789	89	350	68

There is not a significant difference identified in response outcomes overall by original or IP4 refreshment sample classification. Households who had responded at IP6 were, not surprisingly, more likely to respond at IP7. IP4 refreshment sample households that did not respond in IP6 were somewhat more likely to respond in IP7 than original sample households that did not respond in IP6. Similarly, non-responding IP4 refreshment sample households were less likely to refuse at IP7 than the corresponding households from the original sample (noting the small numbers of households). Otherwise, household outcomes were similar across samples regardless of previous wave outcome.

Table 4 below presents household response rates across the two mode conditions: CAPI-only (F2F), and the mixed-mode sequential web-CAPI design (MM). Again, this is only pertinent to the original and IP4 refreshment samples. Total response rate is also broken down into complete (all household members) versus partial (some, but not all, household members) response.

Table 4. Household response at IP7 by CAPI or Mixed-Mode Design

	Total		Original Sample		IP4 Refreshment Sample	
	F2F	MM	F2F	MM	F2F	MM
Responding	74.9% 337	79.4% 680	74.3% 223	80.9% 470	75.2% 115	79.3% 244
Complete HH	52.9% 238	61.0% 516	51.1% 153	61.8% 359	55.6% 85	59.3% 157
Partial HH	22.0% 99	19.4% 164	23.2% 70	19.1% 111	19.6% 30	20.0% 53
Non-contact	11.8% 53	7.9% 67	11.1% 33	7.9% 46	13.1% 20	7.9% 21
Refusals	12.4% 56	9.8% 83	13.1% 39	9.1% 53	11.1% 17	11.3% 30
Other non-responding	0.9% 4	1.2% 16	1.0% 3	2.1% 12	0.7% 1	1.5% 4
Total	450	846	297	581	153	265

The mixed-mode design achieved a significantly higher response rate overall at IP7 (79.4%) than the CAPI-only design (72.9%) (combining complete and partial response). This finding differs from previous waves, where the CAPI-only design had a higher response rate in IP5 while there were no differences between the two designs in IP6. At IP7, this difference is largely driven by the complete households, while there is a smaller difference across designs in partially complete households. The lower response rates also means that non-contact and refusals are higher in the CAPI-only design than the mixed-mode design. These differences are largely the same across both the original sample and IP4 refreshment sample. These findings may in part be due to differences in the incentive structure for mixed-mode and CAPI-only designs (see section 4j).

Turning from the household to the individual, Table 5 presents individual re-interview rates for the original and IP4 refreshment samples. Individuals in households where no contact was achieved are classified as non-contacts, while those in refusing households are classified as individual refusals. Table 6 presents the individual response rate for the IP7 refreshment sample conditional on household response. There were 1680 individual respondents aged 16 or older fully interviewed in IP7 that had been interviewed at a previous wave. As with

household response, there is not a significant difference between original and IP4 refreshment samples. The IP7 refreshment sample added 657 personal interviews (72.2% of eligible enumerated individuals) and 31 proxy interviews (3.4% of eligible enumerated individuals).

Table 5. Individual re-interview response at IP6

	Original Sample	IP4 Refreshment Sample	Total
Personal Interview	65.0% 1137	64.4% 543	64.8% 1680
Proxy Interview	1.8% 31	1.7% 14	1.7% 45
Non-contact	8.2% 144	9.3% 19.1	8.6% 222
Refusal	17.7% 309	19.1% 161	18.1% 470
Other non-response	7.3% 128	5.6% 47	6.8% 177
Total	1749	843	2594

Table 6. Conditional individual response, IP7 refreshment sample

	IP7 Refreshment Sample
Personal Interview	72.2% 657
Proxy Interview	3.4% 31
Non-contact	0.8% 7
Refusal	16.6% 151
Other non-response	7.0% 64
Total	910

The individual-level response rates for the original and IP4 refreshment samples in IP7 across survey mode designs are shown in Table 7 below. The IP7 refreshment sample interviews were only collected by the CAPI-only design. Overall, there are few differences between samples, except a somewhat higher non-contact rate for CAPI-only design in the IP4

refreshment sample. The mixed-mode design has somewhat higher individual re-interview rates, similar to findings from IP6 but contrary to what occurred at IP5.[6] Refusals are higher in the CAPI-only relative to the mixed-mode design; this is the reverse of what occurred in IP6, where the mixed-mode design led to higher individual refusals. The percentage of other non-response is significantly higher for the mixed-mode design across both samples.

Table 7. Individual re-interview response at IP7 by mode

	Total		Original Sample		IP4 Refreshment Sample	
	F2F	MM	F2F	MM	F2F	MM
Personal Interview	61.4%	66.6%	61.9%	66.6%	60.5%	66.7%
	554	1126	364	773	190	353
Proxy Interview	2.0%	1.6%	2.2%	1.6%	1.6%	1.7%
	18	27	13	18	5	9
Non-contact	10.8%	7.4%	9.0%	7.8%	14.0%	6.4%
	97	125	53	91	44	34
Refusal	22.3%	15.9%	23.0%	15.0%	21.0%	18.0%
	201	269	81	174	66	491
Other non-response	3.6%	8.5%	3.9%	9.0%	2.9%	7.2%
	32	143	23	105	9	38
Total	902	1690	588	1161	314	529

At IP6 a "mop-up" phase was introduced where respondents were contacted by telephone to complete the survey, as well as opening the web version to anyone not yet responding. This phase also occurred at IP7 as well for original and IP4 refreshment samples. The IP7 refreshment sample could only complete the survey via CAPI. Table 8 presents the mode actually responded to for all original and IP4 refreshment sample respondents. Not surprisingly, almost all of the CAPI-only assigned respondents completed survey in a face-to-face setting. While the majority of respondents assigned to the mixed-mode design completed the web version, a sizable minority responded when an interviewer approached them at home. However, among those assigned to the mixed-mode design, significantly more respondents in the refreshment sample responded to the web version than those in the original sample, replicating a similar outcome from IP6. A small number of CAPI-only respondents ended up

[6] IP5 had a different incentive structure than IP6 and IP7, which may explain some of these differences.

responding to the web during the mop-up period, with very few respondents in either design or sample responding via the CATI invitation. However, taken together, it is clear that the mop-up phase added a several respondents who otherwise would have been treated as non-productive outcomes.

Table 8. Survey mode of response

	Total		Original Sample		IP4 Refreshment Sample	
	F2F	MM	F2F	MM	F2F	MM
Face-to-Face	95.3%	35.2%	94.2%	38.6%	97.4%	27.8%
	528	396	343	298	185	98
Web	4.5%	64.6%	5.5%	61.5%	2.6%	71.4%
	25	727	20	475	5	353
Telephone	0.2%	0.3%	0.3%	0.0%	0.0%	0.9%
	1	3	1	0	0	3
Total	554	1126	364	773	190	353

4. Experimentation in IP6

There were a number of experiments carried on IP7 covering both fieldwork procedures and measurement in the questionnaire. There were some new experiments and some which were the longitudinal continuation of experiments carried at previous waves of the IP. This section outlines the experiments carried at IP7; briefly explaining the reasons for carrying them, describing the design of the experiment and giving an indication as to the initial results from early analysis of the data. The analyses in this working paper were based on a preliminary data-set which contained all cases but did not have weights or derived variables. The authors, and proposers of the experiment, of each sub-section below are given in the heading.

a. How Do People Think about Environmental Taxes? (Malcolm Fairbrother)

Social science has an important role to play in providing policymakers and advocates with insights into the sources of public opinion about policies for environmental protection. This includes widespread public scepticism about policy measures—such as taxes on pollution and resource consumption—for which there is strong support among experts. Survey research has

already provided some insights into the public's thinking about environmental degradation, protection, and taxation, but there remains considerable scope for further investigation.

This experiment used different versions of a commonly used survey question about people's support for environmental protection in the form of taxation: "How willing would you be to pay higher taxes in order to protect the environment? Not at all willing, not very willing, fairly willing, or very willing?" The aim in methodological terms was to better understand the sensitivity of people's answers to questions about environmental policy. Substantively, the experiment was designed to illuminate the determinants of people's opinions about environmental protection generally. The experiment entailed the random allocation of respondents to one of ten different versions of the question.

Given the content of the ten variants, there were effectively five experiments running simultaneously. These investigated the impact on people's responses of:

1. stating that new environmental taxes would be offset by cuts to other taxes;
2. stating that new revenues from environmental taxes would be spent on the environment (in unspecified ways);
3. drawing respondents' attention to the possibility of the government not doing what it says (having only "promised" to spend the tax revenue on the environment, and perhaps to offset the new taxes);
4. emphasising that the respondents themselves contribute to pollution, through their consumption, with the implication that environmental taxes would affect the cost of "things you buy";
5. making (4) more real by pointing out specific products on which taxes could be raised, and hinting that new environmental taxes might raise the price of these products in particular, which are already heavily taxed in the UK (petrol and electricity).

Of the 2423 respondents included in Wave 7 of the Innovation Panel, 2234 provided valid responses to this question (there were 85 proxies, 11 inapplicable, 6 missing, 4 refusals, and

83 don't knows). So the proportion of respondents who got the question but were unable or unwilling to answer was low (less than 5%).

Table 9 below presents the results of the five experiments, with the raw percentages of respondents who provided each of the four possible answers, under ten different combinations of conditions. The modal answer among respondents who received the basic version of the question (A) was "fairly willing," though substantial numbers selected each possible option. The results for version B show that people are much more willing to pay if new environmental taxes are offset with tax cuts elsewhere—the percentage differences compared to A are large. People are somewhat *less* willing to pay new environmental taxes, however, if told that the revenues will also be spent specifically on programmes for environmental protection (C). Making the possibility of increased taxation more concrete to respondents, and framing respondents themselves as polluters (G), makes no significant difference. But mentioning that the taxes could apply to petrol and electricity specifically reduces support somewhat (I).

The most interesting result comes from comparing the outcome under scenario F with that under scenario D. A significant number of respondents are less supportive of environmental taxes if offsetting tax cuts are presented as only a government "promise" rather than a fait accompli. That is, while cost-neutrality makes environmental protection much more appealing to the public, political distrust appears to reduce the positive effects of framing new environmental taxes as cost-neutral to taxpayers. Considering the relatively minor difference in the wording between versions D and F, the magnitude of the impact on the responses is surprisingly large.

Table 9: Support For Environmental Taxes by Wording Conditions

	Treatments					Responses (%)			
	Offset *	**Spent ***	**Promised †**	**Things You Buy**	**Petrol and Electricity ***	**Not at all willing**	**Not very willing**	**Fairly willing**	**Very willing**
A						29.1	32.9	32.9	5.1
B	X					6.8	18.6	44.3	30.3
C		X				25.0	35.7	35.2	4.1
D	X	X				13.5	20.7	43.7	22.1
E		X	X			27.1	29.4	37.4	6.1
F	X	X	X			22.8	24.2	38.4	14.6
G				X		25.1	30.1	38.9	5.9
H	X			X		9.7	16.1	43.8	30.4
I				X	X	27.7	43.8	25.1	3.4
J	X			X	X	16.0	20.3	42.6	21.1

Raw percentages of respondents giving each answer, depending on the combination of treatments they were assigned.
* effect statistically significant at the 0.05 level (one-tailed)
† effect statistically significant at the 0.05 level (one-tailed), in interaction with Offset

In sum, the experiment generated a number of substantively fascinating findings, and shows the scope for future experiments along these lines.

b. The Impact of Response Scale Direction on Survey Responses (Ting Yan and Florian Keusch)

This experiment manipulates the direction of response scales while holding constant other scale features. The purpose of the experiment is to examine whether and how the direction of a response scale affects survey responses. Four sets of items employing three different scales are subject to this experiment manipulation, as shown in Table 10. Respondents are randomly assigned to descending scales that start with the positive/high end (e.g., strongly agree, completely satisfied, excellent) or ascending scales beginning with the negative/low end (such as strongly disagree, completely dissatisfied, or poor).

Table 10. Survey Items Included in Scale Direction Experiment

Survey Items	Response Scales
1 item measuring general health	Condition 1: poor, fair, good, very good, excellent Condition 2: excellent, very good, good, fair, poor
5 items measuring satisfaction with job, health, income, leisure time, and overall satisfaction	Condition 1: completely dissatisfied, mostly dissatisfied, somewhat dissatisfied, neither satisfied nor dissatisfied, somewhat satisfied, mostly satisfied, completely satisfied Condition 2: completely satisfied, mostly satisfied, somewhat satisfied, neither satisfied nor dissatisfied, somewhat dissatisfied, mostly dissatisfied, completely dissatisfied
20 BIDR items	Condition 1: strongly disagree, somewhat disagree, neither agree nor disagree, somewhat agree, strongly agree Condition 2: strongly agree, somewhat agree, neither agree nor disagree, somewhat disagree, strongly disagree

Preliminary analyses indicate that the scale direction significantly affects resultant survey responses by pushing answers to the beginning of the scale. For instance, when the response scale starts with the negative end ("poor"), 24.0% of respondents reported from the negative side (choosing "poor" or "fair"). However, when the same scale starts with the positive end ("excellent"), the proportion of respondents selecting from the negative side dropped to 16.9%. This difference of 7 percentage points is statistically significant at the 0.05 level, confirming the presence of scale direction effect (see Table 11). In addition, we found that the mode of data collection does not interact with scale direction – scale direction effect is shown in both the CAPI and the Web mode. Furthermore, the significant effect of scale direction still holds after controlling for modes and sample composition (a negative coefficient indicates that scales starting with the positive end reduces the likelihood of choosing from the negative end of the scales).

Similar trends have been found in answers to the 5 satisfaction items. Significantly more reports of dissatisfaction are found when the satisfaction scale starts with dissatisfaction than when the scale starts with satisfaction. As shown in Table 11, the difference in the proportions of respondents reporting from the dissatisfaction side of the satisfaction scale across scale direction is statistically significant for all 5 satisfaction items in both modes of

data collection and the main effect of scale direction is statistically significant after controlling for sample composition and the mode of data collection.

The 20-BIDR items are analyzed together. As shown in Table 11, more people chose from the negative side of scale ("strongly disagree" or "somewhat disagree") when the agreement scale runs from negative to positive than vice versa.

Table 11. Scale Direction Effects Across Items

	Proportion Choosing From the Negative Side of Scale				**Multivariate Model Results**	
	Ascending Scale Negative/Low End Start	Descending Scale Positive/High End Start	Diff.	p-val	Direction Main Effect (Log Scale)	p-val
General Health	24.0%	16.9%	7.1%	<0.0001	-0.48	<0.0001
Job Satisfaction	24.0%	15.7%	8.3%	0.0002	-0.54	0.0002
Health Satisfaction	25.3%	16.2%	9.1%	<0.0001	-0.56	<0.0001
Income Satisfaction	26.3%	20.1%	6.2%	0.0006	-0.34	0.001
Leisure Satisfaction	25.7%	18.2%	7.5%	<0.0001	-0.45	<0.0001
Overall Satisfaction	15.4%	10.1%	5.4%	0.0002	-0.47	0.001
BIDR Items	31.0%	29.4%	1.6%	0.0002	-0.08	0.0003

Our preliminary results demonstrate that scale direction affects survey answers by pushing answers to the start of the scale. There is some evidence suggesting that scale direction effects are more pronounced for longer scales and less pronounced for agreement scales. These results have a great implication for survey researchers and the survey community, once more calling for attention to potential bias induced by scale direction. Further analyses will be conducted to examine the impact of scale direction on relationships. For instance, we will test whether scale direction affects factor patterns for the BIDR items.

c. Separating systematic measurement error components using MTMM in longitudinal studies (Alexandru Cernat and Daniel Oberski)

Measurement error is a pervasive problem in survey research and can cause results to be severely biased. As is well known, attenuation of relationship estimates such as correlations and subgroup differences occurs when non-substantive variations in respondents' answers are unrelated to other answers – i.e. when the errors are random (Lord & Novick, 1968). Systematic measurement error can also occur, however, and can both attenuate and artificially increase apparent relationships (Andrews, 1984; Campbell & Fiske, 1959).

Past research has sometimes found large effects of these systematic components (e.g., McClendon, 1991). It is not clear, however, what the relative importance of each factor might be. This has direct implications for survey question design. For instance, a question designer may note that acquiescence may pose a problem for agree-disagree questions (Hui & Triandis, 1985; Krosnick, 1991) and therefore choose to ask questions directly instead. However, if the direct method were to engender more method effects (e.g., McClendon, 1991), social desirability variance, or extreme response styles, the results could well be even more biased than they would have been with agree-disagree questions. It is therefore essential to estimate not just one type of systematic error effects, but also any trade-offs that might exist with other types of errors.

Design

The design used in this experiment can be described as an extended split-ballot multitrait-multimethod (MTMM) design. In the classical MTMM design (Campbell & Fiske, 1959), multiple traits are evaluated by multiple methods. Andrews (1984) suggested identifying "traits" with "survey items" and "methods" with "item wordings". In what follows we will refer to the combination of a survey item and a particular wording as a "survey question".

It is not difficult to see how the MTMM design may be extended to include other factors besides trait and method effects. In particular, we are interested in separating not only method variance from trait and residual variance, but also acquiescence and social desirability variance. We therefore manipulated the following experimental factors:

- Number of scale points (method): 2 point or 11 point scale

- Socially desirable direction: positively or negatively formulated item on immigration
- Acquiescence direction: Agree-disagree or Disagree-agree scale

In order to estimate the model of interest we ask 6 questions about immigration attitudes (see Appendix) at the beginning of the questionnaire, and 6 questions on the same topic in another wording format at the end, after 20 minutes or more of other questions are asked. The combination of the factors manipulated leads to 28 possible wordings of the questions. By also randomizing the ordering of the wordings, we have created 56 experimental groups.

Initial results

Due to the novelty of the design and the modelling the first step of the research is finding the measurement model that fits the data. Initial analyses have led to a model that includes only five of the traits (T2-T6), an overall factor for attitudes towards immigrants, a social desirability factor and an acquiescence factor. The method factors are not modelled here in order to decrease computational time.

To evaluate if the latent variables measure what we affirm, we correlate them with a number of substantial variables. The expected relationships are based on previous research and theoretical support:

Substantive traits (T2-T6 and overall):

- Political party supported (left-right scale);
- Unemployed (threat theory);
- Non-British identity (self-interest/experience).

Social Desirability factor:

- "Balanced Inventory of Desirable Responding" (BIDR_1 and BIDR_2);
- Whether anybody else was present during the interview;
- Whether you admit to having taken drugs.

Acquiescence factor:

- An Acquiescence Response Style (ARS) measure constructed from different questions.

BIDR_2, the social desirability scale, is not related to anything else besides the social desirability factor. The latter is, in turn, also related to the presence of others as well as admitting to drug use. The substantive factors, Trait 2-5 and overall, are related to party support and being non-British but not to being unemployed. Acquiescence is not related to the two ARS measures calculated from other questions.

Thus, these findings partially support our theoretical expectations regarding the latent variables. Further research is needed to develop the full model and to validate the latent factors.

Table 12: Correlations between factors and criterion variables

	Overall	**Race (T2)**	**Poor (T3)**	**Economy (T4)**	**Enriched (T5)**	**Better place (T6)**	**Social Des.**	**Acquiescence**
Party support	*	*	*	*	*	*	*	*
Unemployed								
Non British	*	*	*	*	*	*	*	*
Non UK born							*	
BIDR 1[+]								
BIDR 2							*	
Int. present	*	*	*	*		*	*	
Drugs	*	*	*				*	
Age	*	*	*		*	*	*	
ARS 1[++]								
ARS 2					*			

[+] ***Balanced inventory of desirable responding;*** [++] ***Acquiescence Response Style from different measures***
**** Indicates a significant correlation***

d. Experiments for survey question evaluation in multiple-country contexts (Henning Silber, Jon A. Krosnick, Tobias H. Stark, Annelies G. Blom, and Peter Lynn)

Our research explores whether the principles of question design afloat in the field today, which are based on American data, can legitimately be generalized across countries. More precisely, our research implemented well-tested split-ballot design experiments from single-country contexts in multiple countries to gauge country-specific differences in response behaviour and satisficing. This report describes the results of the experiments with the Innovation Panel.

We replicated seventeen split ballot experiments from Schuman and Presser (1981) and Schuman and Ludwig (1983). The experiments tested for differences in response behaviour by altering the order in which (a) the response options and (b) the questions were presented. Moreover, we tested (c) for differences caused by acquiescence (a tendency to agree with any presented statement), (d) for effects of different no opinion filters, (e) for question balance effects (balanced questions are completely neutral), and (f) for differences caused by the tone of wording of a question. Each of the seventeen experiments involved administering two versions of a question, each to a random half of the survey respondents. We expected the classic effects to replicate in the UK.

Results

Since the survey mode has a strong impact on question evaluation and on designing good questions, we present the results separately for the CAPI and CAWI mode of response. Schuman and Presser's (1981) original data collections were all done with oral administration, so they are closest to the CAPI mode.

We present here results from two of the seventeen experiments, plus a summary overview of the results of the remaining experiments.

Response Order Experiment: Oil Supply

We replicated an experiment of Schuman and Presser (1981) that involved comparing the following two questions:

Form A: *Some people say that we will have plenty of oil 25 years from now. Others say that at the rate we are using our oil, it will all be used up in 15 years. Which of these ideas would you guess is most nearly right?* ("There will be plenty of oil in 25 years", "Oil will be used up in 15 years")

Form B: *Some people say that at the rate we are using our oil, it will be all used up in 15 years. Others say that we will still have plenty of oil 25 years from now. Which of these ideas would you guess is most nearly right?* (Categories: "Oil will be used up in 15 years", "There will be plenty of oil in 25 years")

For this experiment, Schuman and Presser (1981) reported a significant recency effect (effect size = 13.8 percent, p-value = .00). Table 13 and Table 14 show the results separately for the CAPI and CAWI respondents. The CAPI mode showed the expected significant recency effect (effect size = 10.6 percent, p-value = .01). Respondents were significantly more likely to say "There will be plenty of oil in 25 years" when that response option was presented last. The CAWI mode showed a non-significant effect in the expected direction (effect size = 3.7 percent, p-value = .48). The difference between the modes was non-significant (response X mode X form: $\chi 2 = 1.16$, df = 1, p-value = .28).

Table 13: Response Order Experiment on Oil Supply – CAPI

Response	Plenty First	Used Up First	Difference	$\chi 2$	df	p
Plenty	48.7%	59.3%	10.6%	7.86	1	.01
Used Up	51.3%	40.7%	-10.6%			
Total	100%	100%				
N	355	344				

Table 14: Response Order Experiment on Oil Supply – CAWI

Response	Plenty First	Used Up First	Difference	$\chi 2$	df	P
Plenty	49.7%	53.4%	3.7%	0.49	1	.48
Used Up	50.3%	46.6%	-3.7%			
Total	100%	100%				
N	187	176				

* light blue = expected direction and significant, ** light green = expected direction and non-significant, *** pink = opposite direction and non-significant, **** dark red = opposite direction and non-significant

Acquiescence Experiment: Individuals vs. Social Conditions

Another experiment of Schuman and Presser (1981) involved comparing responses to these two questions:

Form A – Forced Choice (FC): *Which in your opinion is more to blame for crime and lawlessness in this country – individuals or social conditions?* (Categories: "Individuals more to blame", "Social conditions more to blame")

Form B: *Do you agree or disagree with this statement? Individuals are more to blame than social conditions for crime and lawlessness in this country.* (Categories: "Agree", "Disagree")

Schuman and Presser (1981) reported a significant agreement effect (effect size = 13.2 percent, p-value = .00). Table 15 and Table 16 show the results separately for CAPI and CAWI respondents. The CAPI mode showed the expected significant agreement effect (effect size = 16.1 percent, p-value = .00). Respondents were significantly more likely to agree with the statement when the question was asked in the agree/disagree format. The CAWI mode too showed the expected significant agreement effect (effect size = 12.5 percent, p-value = .01). The difference between the modes was non-significant (response X mode X form: $\chi 2$ = .21, df = 1, p-value = .65).

Table 15: Acquiescence Experiment on I vs. SC- CAPI

Response	FC (I)	Agree (I)	Difference	χ2	df	p
Individuals	56.6%	72.7%	16.1%	21.18	1	.00
Social Conditions	43.4%	27.3%	-16.1%			
Total	100%	100%				
N	394	362				

Table 16: Acquiescence Experiment on I vs. SC- CAWI

Response	FC (I)	Agree (I)	Difference	χ2	df	p
Individuals	61.9%	74.4%	12.5%	6.17	1	.01
Social Conditions	38.1%	25.6%	-12.5%			
Total	100%	100%				
N	181	164				

Conclusion

Overall, in the CAPI mode twelve of the seventeen experiments replicated the findings of the original investigators, four showed non-significant effects in the expected direction, one showed a non-significant effect in the opposite direction, and none showed a significant effect in the opposite direction (see Table 17). In the CAWI mode eight of the seventeen experiments replicated, six showed non-significant effects in the expected direction, three showed non-significant effects in the opposite direction, and none showed a significant effect in the opposite direction (see Table 18)

Table 17: Summary of the Experimental Results - CAPI

Types of Experiments	Number of Experiments				
	Replicated and significant	Replicated and non-significant	Opposite direction and non-significant	Opposite direction and significant	Total
Response Order	4	1			5
Acquiescence	2	1	1		4
Question Wording		1			1
No Opinion	3				3
Question Balance	2				2
Question Order	1	1			2
Total	12	4	1	0	17

Table 18: Summary of the Experimental Results - CAWI

Types of Experiments	Number of Experiments				
	Replicated and significant	Replicated and non-significant	Opposite direction and non-significant	Opposite direction and significant	Total
Response Order	1	3	1		5
Acquiescence	2	1	1		4
Question Wording			1		1
No Opinion	3				3
Question Balance	1	1			2
Question Order	1	1			2
Total	8	6	3	0	17

As expected, many of the experiments of the classic experiments replicated. Specifically, the experiments with the no opinion response option, the experiments with the counterargument, and the question order experiments showed only effects in the expected directions. However, the results also showed some remarkable differences in comparison with the classic results. Specifically, a question wording experiment with neutral and liberal versions of a question about free speech did not replicate the differences found by Schuman and Presser (1981) when these two versions were compared:

> Form A: *There are some people who are against all churches and religion. If such a person wanted to make a speech in your city/town/community against churches and religion, should he be allowed the freedom to speak, or not?* (Categories: "Yes, allowed to speak", "No, not allowed")
>
> Form B: *There are some people who are against all churches and religion. If such a person wanted to make a speech in your city/town/community against churches and religion, should he be allowed to speak, or not?* (Categories: "Yes, allowed to speak", "No, not allowed")

Additionally, the experiments showed some differences in results between the CAPI and CAWI survey modes. Particularly, response order experiments and a question order experiment on "spending" showed different effects. Possible explanations for these mode differences are the different presentation formats (oral in CAPI and visual in CAWI) and the social presence of an interviewer in the CAPI mode. The comparison of these results with other countries will allow us to explore whether the results are culture-specific for the United Kingdom and the United States or can be generalized across other countries as well.

e. Use of multiple contacts between waves (Jonathan Burton)

The use of inter-wave mailings is standard on longitudinal studies, but is there an added benefit of multiple contacts between waves?

1 Do multiple contacts reduce the proportion of the sample who are not contacted or move house and are unable to be traced?

2 Do multiple contacts reduce the proportion of the sample who refuse to participate in the study?

Research has shown that keeping track of sample members on a longitudinal study between waves helps minimise attrition (Couper and Ofstedal 2009; Laurie et al. 1999). A household, or individual, which is found to have moved when an interviewer calls requires tracing attempts which are time consuming and are not always successful. Losing a sample member because they cannot be traced is a regrettable outcome, since if they are found they may be quite willing to continue participation. In addition, by under-representing those who move, we are likely to be under-representing a number of important life outcomes and thus under-count measures of change. Therefore, longitudinal studies use a number of mechanisms to keep in contact with sample members, and to make tracing easier if they are found to have moved. These mechanisms include change-of-address cards, collecting email address, mobile/work telephone numbers, collecting contact details of a stable contact and sending a report of findings between waves.

There has been research which has looked at the style of between-wave mailing, and the number of mailings – but these have tended to be with surveys with longer intervals between waves than one year. For example, research in the US (McGonagle, Couper and Schoeni, 2011) on the biennial (since 1997) PSID suggests that multiple contact attempts can increase the proportion who supply updated information, but had no effect on whether the household required tracking.

On the main-stage of *Understanding Society*, the practice of one inter-wave mailing to each adult per year has changed since 2013, following qualitative research by NatCen (unpublished) which recommended more frequent contact with sample members. Under the new strategy, there are multiple contacts throughout the year (two short reports and one more substantial report). This strategy was implemented across the whole sample on the main-stage of *Understanding Society*, but this experiment gives us the chance to measure the effect of multiple contacts, compared to a single contact between waves.

<u>What are the benefits of multiple mailings?</u>

On *Understanding Society*, as other longitudinal panel studies, the main causes of attrition are failure to locate a household, failure to contact a household once located and failure to persuade a sample member to participate when contacted (Lepkowski and Couper, 2002).

Failure to locate: This is where an interviewer finds that a household containing sample members has changed their address, and cannot get the new address. On *Understanding Society*, this would be classified as an untraced mover. The study issues sample members with a change-of-address card with their advance letter, and with the inter-wave mailing. Increasing the number of inter-wave mailings, increases the number of change-of-address cards the sample member receives between waves. It is hoped that this would encourage sample members to notify us of any change in their address, with the expectation that multiple mailings would mean that there is a greater chance that the sample member received a change-of-address card close to their date of residential move. Where an inter-wave mailing is returned to us by the Post Office as the addressee is "no longer at the address", this gives us a chance to start tracing the sample member before the survey goes into the field.

Failure to contact: This is where the interviewer has attempted to contact the sample member but has not been successful. This may be because the interviewer never calls when the sample member is at home – or it may be that the household has moved from the address, but the interviewer never contacts anyone to identify that this has happened. In the latter case, the use of multiple contacts during the year would act to reduce the level of untraced movers, as above. In the former case, we would not expect multiple mailings to affect the sample members' propensity to be at home.

Failure to persuade: Once an interviewer has located and contacted a sample member, they may still not be successful in getting an interview because the sample member may refuse to participate. The hope is that the use of multiple mailings will reduce the propensity to refuse, by increasing the study's profile with the sample member, increasing recognition of the name, logo and 'brand', and the sense of 'belonging' to the study. Additional mailings give us a chance to give the sample member more information about how the study is used, and make them feel more valuable and encourage them to see the value in the study and their continued participation. Different styles and content of the inter-wave mailing give us more opportunity to appeal to different types of sample member.

Potential risks of multiple mailings

There may also be some risks to using multiple mailings between waves. It is possible that increased contacts will irritate sample members, who may start to see the contacts as 'junk

mail' or as a waste of resources, which may lead to higher refusal. Additional contact attempts also provide additional opportunities for sample members to contact us to refuse further participation.

IP7 experiment

This experiment gauges the effect of multiple contacts on (i) locating, (ii) contacting and (iii) responding at IP7. Adults in half of households were randomly allocated to receiving one between-wave mailing between IP6 and IP7, the adults in the other half of households received three mailings.[7]

IP6 finished fieldwork on the 16th July, 2013. The first deliveries of data from the fieldwork agency were in early October. This meant that the earliest an inter-wave mailing file could be generated was mid-October. Table 19 below indicates the inter-wave mailing content received by each treatment group. The footnotes link to the Understanding Society web-site which includes more information on the subject covered and has a picture of the front of the mailing that was sent. It should be noted that inter-wave mailings are sent to each adult in the household, rather than one mailing to the whole household.

Table 19: Timing and content of IP6-to-IP7 inter-wave mailings

Mailing date	Single mailing (control)	Multiple mailings
25 October 2013	None	Poverty[8] or well-being[9]
10 December 2013	"The story so far"[10]	"The story so far"[4]
11 February 2014	None	Media headlines[11]

Results

The outcome of interest in this experiment is the interview outcome at IP7. If multiple between-wave mailings encourage sample members to keep in touch with us and inform us of

[7] At IP7 there was a new refreshment sample, this sample did not receive any inter-wave mailings and so are not included in these analyses.
[8] https://www.understandingsociety.ac.uk/participants/features/poverty-time-to-take-a-fresh-look
[9] https://www.understandingsociety.ac.uk/participants/features/developing-a-wider-picture-of-well-being
[10] https://www.understandingsociety.ac.uk/participants/features/insights-2013
[11] https://www.understandingsociety.ac.uk/participants/features/headlines

their changes of address, we should see lower rates of non-contact and untraced movers. If the effect of multiple mailings is to inform and motivate sample members to participate because they are more engaged with the study, we should see lower rates of refusal.

Table 20, below, shows the household outcome at IP7 for the whole eligible sample, and split by sample group. As can be seen, there is almost no difference in response rates between the treatment and control groups. There are no statistically significant differences for any outcome between the groups.

Table 20: Household outcomes by sample, single and multiple inter-wave mailings

	Whole sample		IP1 original		IP4 refreshment	
	Single	Multiple	Single	Multiple	Single	Multiple
	%	%	%	%	%	%
Responding	77.7	78.0	77.0	80.0	79.3	74.0
Refusal	12.4	12.0	12.4	11.3	12.3	13.5
Non-contact	3.7	3.1	3.5	2.4	3.9	4.7
Untraced mover	3.7	5.0	4.7	4.5	1.5	6.1
Other non-response	2.6	1.9	2.4	1.9	3.0	1.9
Total n	655	640	452	425	203	215

Turning to an individual-level analysis, Table 21 below shows that there is no impact of multiple between-wave mailings for individuals. Similarly to the household-level analysis, those in the IP4 refreshment sample seem to be more likely to respond after receiving the single mailing, and the original IP1 sample more likely after the multiple mailings. However, the only statistically significant difference is for those in the original IP1 sample where those who received multiple mailings were more likely to respond.

Table 21: Individual-level outcomes, single and multiple inter-wave mailings

	Whole sample		IP1 original		IP4 refreshment	
	Single	Multiple	Single	Multiple	Single	Multiple
	%	%	%	%	%	%
Responding	66.6	68.7	66.1	70.2*	67.6	65.7
Refusal	10.9	10.0	11.3	9.9	10.2	10.0
Non-contact	4.2	3.1	4.6	2.2	3.2	4.9
Other non-response	1.2	1.2	1.4	1.4	0.7	0.9
Non-contact HH	6.6	6.3	6.5	5.4	6.9	8.0
Refusal HH	9.2	9.2	8.9	8.7	10.0	10.2
Other NR HH	1.3	1.6	1.2	2.3	1.4	0.2
Total n	1409	1325	977	876	432	449

* significant at $p<0.05$

The impact of multiple mailings had no effect on response at IP7 for those adults who had said at IP6 that they wanted to move or expected to move house (not shown). The use of multiple mailings also had no significant effect on the number of calls required by the interviewer to complete the household: a mean of 4.6 calls for households who received multiple mailings compared to a mean of 4.7 calls for those who received a single mailing.

Conclusion

It appears that using multiple mailings between waves did **not** significantly reduce the proportion of untraced movers, non-contacts or refusals, compared with those who only received a single mailing. This is a surprise finding, given our expectations. It may be that sample members do not read the inter-wave reports and may simply throw out the findings and change of address card. We occasionally hear at interviewer briefings that sample members do not remember receiving any report between waves. In a short online follow-up survey in April 2014 we asked IP sample members for whom we had an email address to answer a few questions. The response rate was low, with 18.7% of those who had eligible

email addresses responding.[12] We can only draw some indicative findings from this short survey, but almost nine-in-ten reported reading the inter-wave mailings. However, when we asked people which reports they remembered receiving (with a picture of the report), those who had only received a single mailing reporting seeing those which they did not receive.

It should be noted that these multiple mailings were over a relatively short period of time – less than four months and the first one was sent some time after the interviews. The late-October mailing was more than six months after the earliest interviews on IP6. Those households or individuals who were going to move house between IP6 and IP7 thus had half of the year where they could have moved before receiving a between-wave mailing with a change-of-address card. It may be that multiple mailings would be more effective if the first was received sooner after the interview than in this experiment.

Future work

On the main survey, the lag between interview and inter-wave mailing has been reduced, through the use of early deliveries of data from the fieldwork agencies. This means that those sample members issued to field in January to March (and could be interviewed at any point from January to September) are mailed in June. This requires additional effort and resource, but it means that individuals are likely to receive a mailing sooner after their interview.

During the period that IP7 was in the field the work on refreshing the participant engagement and communication materials has developed further. Rather than just postcards, the sample members now receive more interesting formats of materials, and the design of these mailings has been improved and is tied in to additional information available on the Participant website. More recent mailings are available to view on-line.[13] Future research may need to focus not on the frequency, but the content of mailings. Creating a balance between easy-to-read and accessible, but still conveying the depth and breadth of the research is a challenge that those who manage longitudinal surveys try to meet – for Understanding Society, four times a year.

[12] An email invitation was sent to 1,435 email addresses, of which 105 bounced. Of the remaining 1330 emails received, 249 sample members started the short survey.
[13] https://www.understandingsociety.ac.uk/participants/features

f. The use of conditional and unconditional incentives (Jonathan Burton)

Research on the use of incentives generally finds that unconditional incentives are more effective than conditional incentives (Church 1993; Goyder 1994; Hopkins and Gullickson 1992; Singer et al. 1999). However, on a longitudinal study, after the first couple of waves, it may be that sending unconditional incentives to previous non-responding individuals within responding households is a waste of resources, which could be more effectively re-directed elsewhere. The 'rewarding' of serial non-responders may also weaken the motivation of responders within the household to participate.

There has been some research on longitudinal studies which has found conditional incentives performing better than unconditional incentives (Castiglioni, Pforr and Krieger (2008) with a CAPI survey) and also the reverse finding (Jäckle and Lynn (2008) on a mail survey). Anecdotal evidence from interviewers suggests that they would prefer to have the incentive in their hands to use to persuade people on the door-step. This may be particularly effective for those who have not participated before (and where the unconditional incentive has therefore been ineffective in the past).

The Innovation Panel has experimented in the past with the size of the incentive, including groups where there is a conditional element – although these have been in addition to an unconditional element. The IP has not tested the effect of conditional experiments on a longitudinal study.

This experiment randomly allocated households to two groups. In one the incentive was sent unconditionally to every adult as usual, including those who had not participated at IP6. In the other, only those adults who participated in the previous wave were sent an unconditional incentive. The advance letter for previous-wave non-responders in this treatment group informed the sample member that the interviewer would give them a voucher if they participate.

<u>Results</u>

There was no significant difference at the individual-level between those who received an unconditional incentive and those who were promised an incentive if they took part. Table 22

below shows that the individual-level response was 22.8% among adults who received their voucher with the advance letter, compared to 27% who had not.

Table 22: Individual-level response rate for those who did not participate at IP6

	Unconditional incentives %	Conditional incentive %
Response	22.8	27.0
Non-contact	7.4	5.2
Refusal	15.9	10.9
Other non-interview	7.4	2.9
Non-interviewed household	46.6	54.0
Total	189	174

Those in the unconditional group were sent their vouchers with the advance letters, and they were able to keep their voucher whether they participated or not. Those in the conditional incentive group were only given vouchers if they participated. Thus, if there is no significant difference in the response rate, using conditional vouchers will be more cost-effective.

At IP7, for this sample, there were three incentive groups; receiving £10, £10 plus £20 if all adults in the household participated, and £30. Table 23 below sets out the cost of the incentives for each group. For the 189 individuals in the unconditional group, all received vouchers. The total cost of incentives was £2530. For 43 respondents, this equates to £58.84 per respondent. In the conditional incentive group, only the respondents received the incentives. The total cost of incentives was £1030, which equates to £21.91 per respondent.

Table 23: The cost of incentives, by treatment group

	Unconditional incentives		Conditional incentives	
	Number in group	**Number received**	**Number in group**	**Number received**
£10 incentive	131	131 x £10	103	18 x £10
£10 + £20 incentive	31	31 x £10	28	7 x £10
+ Eligible for £20	5	5 x £20	6	6 x £20
£30 incentive	27	27 x £30	43	22 x £30
Total incentives	189	£2530	47	£1030
Per respondent	43	£58.84	47	£21.91

Conclusion

Sending out unconditional vouchers is more expensive than promising conditional incentives. However past research has shown that unconditional incentives are also more effective at getting a good response rate. The costs of the unconditional incentives are seen as a price worth paying for a high response rate. However, once a sample has been recruited, and has been interviewed for a number of waves, the act of sending incentives to those who do not regularly participate may not be an effective use of resources. This research has suggested that offering only conditional incentives to previous-wave non-responders has no significant effect on response rate, but the cost per respondent is just over one-third of the cost of sending unconditional incentives to this group. This saving may be diverted into other methods to increase response among this group, for instance by increasing the level of the conditional incentive or incentivising interviewers to make additional efforts.

g. The reliability of measures of change in self-assessed disability (Steve Pudney and Annette Jäckle)

Background

This experiment is a repetition of an experiment carried in IP6. The experiment used reactive dependent interviewing to investigate the measurement of change in self-assessed measures of long-standing illness or disability. For a description of the background and motivation for

this experiment, see the *Understanding Society* Working Paper describing experiments in IP6 (Al Baghal (ed.) 2014).

Objectives

This experiment had three main objectives: (1) to identify the reasons for the high rates of year-on-year change in long-term illness or disability observed at the individual level; (2) to investigate whether use of the initial filter question has a significant impact on measured disability by barring access to the more specific question about everyday activities; (3) consequently, to suggest options for redesigning the questions to give more stable measures.

Experimental design

Sample members were randomly (by household) allocated to one of three experimental groups. At IP7 the experiment was repeated with respondents being allocated to the same treatments as in IP6. The IP7 refreshment sample was allocated to group C and is excluded from the analyses presented here.

Group A (quarter of the sample):

Received the standard version of questions in the general health module, i.e. the HEALTH filter followed by the Activities of Daily Life (ADL) question for respondents who answer "yes" to the filter:

HEALTH: *Do you have any long-standing physical or mental impairment, illness or disability? By 'long-standing' we mean anything that has troubled you over a period of at least 12 months or that is likely to trouble you over a period of at least 12 months. (Yes/No)*

If HEALTH=yes:

ADL: *Does this/Do these health problem(s) or disability(ies) mean that you have substantial difficulties with any of the following areas of your life?*

1 Mobility (moving around at home and walking)

2 Lifting, carrying or moving objects

3 Manual dexterity (using your hands to carry out everyday tasks)

4 Continence (bladder and bowel control)

5 Hearing (apart from using a standard hearing aid)

6 Sight (apart from wearing standard glasses)

7 Communication or speech problems

8 Memory or ability to concentrate, learn or understand

9 Recognising when you are in physical danger

10 Your physical co-ordination (e.g. balance)

11 Difficulties with own personal care (e.g. getting dressed, taking a bath or shower)

12 Other health problem or disability

96 None of these

Group B (quarter of the sample):

Everyone was asked the ADL question; the HEALTH filter question was not asked.

Group C (half the sample):

Everyone was asked the HEALTH question about long-standing health conditions. Respondents who gave a different answer from the previous wave were asked a follow-up question about the reasons for the change:

Can I just check, our records show that last time when we interviewed you on [ff_intdate], {you had a / you did not have any} long-standing illness or disability. Is there an error in our records, or {do you no longer have this condition / is this a new condition}?

Everyone in this group was also asked the ADL question, but at a later point in the questionnaire.

Results

Table 24 documents the explanations respondents in Group C gave for changes in their long-term illness or disability status. In both waves 6 and 7 only few of the respondents who no longer reported having a long term health condition or disability confirmed that it had indeed ended (11 of 45 respondents in wave 6; 5 of 49 respondents in wave 7). The majority of respondents said that they still had the same condition, but that it was not as bad now, or medication/treatment was more effective, or it was less of a problem because their activities

had changed (30 of 45 respondents in wave 6; 38 of 49 respondents in wave 7). Only few respondents (5 in waves 6; 9 in wave 7) said there was an error in their data from the previous interview or other reason for the change in their health status.

Among respondents who reported the onset of a new long-term health condition, in both waves close to half confirmed that this was a new health condition. Measuring the onset seemed more problematic at wave 6 than at wave 7: 25 of 68 respondents said there was an error in the data from the previous wave, or other reason; at wave 7 only 7 of 49 respondents reported a problem with the data or other reason. Some respondents said that they had had the condition previously, but that it had got worse, medication was less effective, or their activities had changed such that it was now more problematic (15 of 68 respondents in wave 6; 8 of 26 respondents in wave 7).

Table 24: Reasons for changes in long-term health status (Group C)

Reasons for no longer reporting long term illness/disability (N)	**Wave 6**	**Wave 7**
There is an error in the records	4	7
I still have the same health condition but it is not as bad now	9	16
I still have the same health condition but treatment or medication is effective now	15	17
The condition is much the same as last year, but my activities have changed, so it is less of a problem now	6	5
I no longer have this health condition	11	5
Other reason	1	2
Total responses (respondents)	46 (45)	52 (49)
Reasons for reporting new long term illness/disability (N)	**Wave 6**	**Wave 7**
There is an error in the records	20	3
I had the same health condition but it is worse now	9	6
I had the same health condition but treatment or medication is less effective now	3	1
The condition is much the same as last year, but my activities have changed, so it is more of a problem now	3	1
This is a new health condition	29	12
Other reason	5	3
Total responses (respondents)	69 (68)	26 (26)

Disability rates based on the questions about Difficulties with Activities of Daily Life tended to be higher if everyone was asked this question, than when respondents were only routed into this question if they reported a long-term illness or disability (Table 25). In wave 6 the rates were 30% versus 24% (P=0.0876); in wave 7 30% versus 20% (P=0.0038). Correspondingly the mean numbers of activities that respondents had difficulty with tended to be somewhat higher if everyone was asked the question than when it was routed (wave 6: 0.64 versus 0.51, P=0.1331; wave 7: 0.71 versus 0.5, P=0.0511).

Table 25: Difficulties with Activities of Daily Life

	% difficulty with 1+ ADL			Mean number of difficulties with ADLs		
	Group B (no filter)	Group A (filter)	P-value	Group B (no filter)	Group A (filter)	P-value
Wave 6	30%	24%	0.088	0.64	0.51	0.133
Wave 7	30%	20%	0.004	0.71	0.50	0.051

Table 26 documents entry and exit rates among respondents who answered the health questions in two consecutive waves. Using the HEALTH question, the entry rates (that is, the proportion of respondents who did not report a long term health condition or disability in wave 6, but did report a condition in wave 7) was 7%. The exit rate (that is the proportion of respondents who reported a condition in wave 6, but not in wave 7) was 25%.

Using the follow-up question about difficulties with Activities of Daily Living (ADL), the entry rate was similar at 6% when the ADL question was filtered and only asked if the respondent reported a long term health condition in response to the HEALTH question (Group A). When the ADL question was not filtered but instead asked of everyone (Group B), the entry rate was higher at 12%. Exit rates where always higher with the ADL questions than the HEALTH question at 37% when ADL was filtered on HEALTH, and 32% when not filtered.

Table 26: Entry and exit rates

Variable used to identify change in health status	Waves (experimental groups)	Entry rate	Exit rate
HEALTH (filter question)	Waves 6-7 (groups A+C)	7%	25%
Activities of Daily Living	Waves 6-7 (group A: filtered)	6%	37%
	Waves 6-7 (group B: unfiltered)	12%	32%

h. Wording effects of dependent interviewing questions on the amount of change observed in panel data (Annette Jäckle, Noah Uhrig, and Emanuela Sala)

Introduction

This study investigates how best to word dependent interviewing questions. *Understanding Society* uses dependent interviewing for many questions. For some items respondents are reminded of their answer in the previous interview and asked whether this is still the case, for other items they are asked whether this has changed. In IP3/IP4 we implemented an experiment contrasting these two versions. The analyses showed clear evidence of agreement bias: respondents were much more likely to say "Yes, this is still the case, than "No, this has not changed" – and much more likely to say "Yes, this has changed" than "No, this is not still the case". In this follow-up study we aimed to contrast the two question formats with a 'balanced' format which does not allow simple confirmation. To study potential response order effects, and whether these differ in CAPI and web, we also varied the order of response options in the new format. Contrasting CAPI and web allows studying differences in agreement bias in the yes/no questions, and differences in response order effects in the 'balanced' format. Note that this is a replication of an experiment originally developed for IP5, which was corrupted by errors in the implementation of experimental treatments.

Design

The experiment was replicated over a total of 19 items in the household and individual questionnaires. In the household questionnaire the items were the number of rooms, housing tenure, amount of mortgage payments and amount of rent payments. In the individual questionnaire the items were wanting to move, type of educational institution, whether job is permanent, industry, occupation, whether employee or self-employed, size of workplace, hours of work (employees and self-employed), gross earnings, net earnings, whether salaried

or paid by the hour, mode of transport to work (employees and self-employed), and whether self-employed work on their own or in partnership. In the face-to-face interviews the items used for this experiment were audio-recorded.

Respondents were randomly allocated to one of four dependent interviewing wordings. That is, the design was a between-respondent design to avoid confusing respondents. In all groups respondents were first reminded of the answer they had given in the previous interview. The groups varied in the wording of the follow-up question used to determine the current status. The four versions of the follow-up question were:

1) Is this still the case? (Yes/No)
2) Has this changed? (Yes/No)
3) Is this still the case or has it changed? (Still the case/Has changed)
4) Has this changed or is it still the case? (Has changed/Is still the case)

The data contain observations on 11,446 dependent interviewing questions, nested in 1,628 respondents. The number of items answered varied between respondents due to routing and because some respondents completed only the individual or household interview, while others completed both.

Analysis methods

To test whether the question wording affected the probability of reporting a change, and whether wording effects were different between modes of interview, we pooled all experimental dependent interviewing questions and derived a binary indicator for whether or not the respondent reported a change.

Figure 1 shows the percentage of respondents who reported a change, by mode of interview and by dependent interviewing question version. To test for significance of differences between question versions and between modes we used two approaches. First we estimated a linear probability model by regressing whether or not the respondent reported a change on a binary indicator for the mode of interview, dummy variables for the DI question wording, and interactions between the mode and the DI dummy variables. Standard errors were adjusted for clustering of items in respondents. This model tests whether question wording affected responses, and whether there were differences in wording effects between those respondents who completed the interview face-to-face and those who completed by web. This method

does however not account for selection into mode. The estimated effect of the mode therefore includes both effects due to differences in the sample composition between modes, and causal effects of the mode on the question wording experiment.

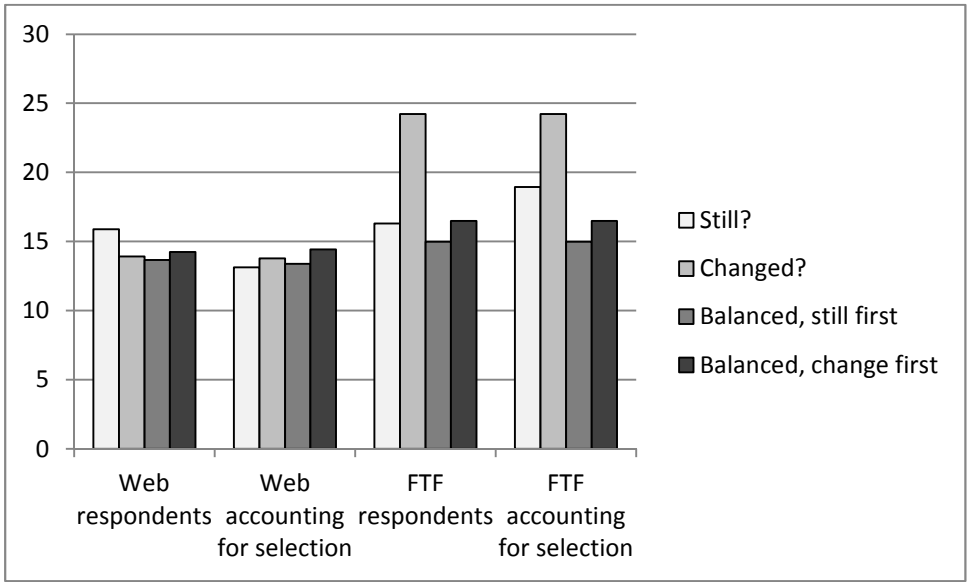

Figure 1: Percent of respondents reporting a change, by mode of interview and DI wording

To account for selection into mode and test for any causal differences between modes in the effects of the question wording, we used an instrumental variable approach as described by Greenland (2000). IP7 included a mixed-mode experiment that was crossed with the dependent interviewing wording experiment. The mixed-mode experiment can be thought of as an "encouragement design": the control group was allocated to face-to-face interviewing only (for this analysis we excluded 27 respondents in the control group who completed the interview by web in a final non-response mop-up stage). The treatment group were invited to complete the survey online. Non-respondents were later followed up by face-to-face interviewers. That is, those who did not comply with the web request reverted to the control treatment. With this setup the initial randomisation into face-to-face only versus sequential web and face-to-face can be used as an instrument to obtain an unbiased estimate of the effect of mode on responses. We estimated a two-stage least squares linear probability model including covariates and adjustment for clustering of items in respondents as above, but using the initial randomisation as an instrument for the mode of interview.

Results

Among web respondents the DI question wording had no effect: depending on the wording between 14% and 16% of respondents reported a change. This null effect was confirmed when controlling for selection into mode, with 13-14% of respondents reporting a change.

Among face-to-face respondents those who were asked “Has that changed?” were significantly more likely to report a change (24%) than respondents who were asked any of the other question versions (15-16% and 15-19% when taking account of selection into face-to-face interviewing).

Comparing responses to the different DI question versions between modes, the “Has that changed?” version was the only wording where the probability of reporting a change was significantly different. Face-to-face respondents were 10 percentage points more likely to report a change than web respondents and this result remained significant after controlling for selection into mode.

The results suggest that the effects of the DI question wording on the reporting of change are mediated by the interaction between interviewers and respondents. Further investigation will include analysing data from the audio-recordings of face-to-face interviews.

i. Social Desirability in General Health Questionnaire (GHQ-12) (Olena Kaminska and Cara Booker)

Sensitive questions are frequently asked on social surveys, and more often than not they are related to important social issues with direct policy implications. These topics include alcohol consumption and smoking, life satisfaction, happiness, health, sleeping problems and a variety of (psychological and sexual) disorders (behaviours) to name just a few. Much has been researched on how to improve survey questions of this type, but most of the efforts focus on reducing potential respondent embarrassment. The theory of embarrassment suggests that a respondent has comprehended the question carefully and as intended, searched the memory, come up with the correct answer, and is aware of it. At the reporting stage the respondent nevertheless gets embarrassed and purposely misreports the answer in order to appear in more favourable way (Tourangeau, Rips and Rasinski, 2000).

While the effect of embarrassment is often present, we argue that this may not be the only cause of social desirability. A recent study by Kaminska and Foulsham (2013) suggest that satisficing may lead to social desirability, especially where response scale is presented vertically and starts with the most positive option at the top. By observing respondents' readings of the questions via eye-tracking, the authors found that 47% of questions do not receive any looks at the bottom response options (which were negatively worded). In addition, respondents who rush through the questions (have higher than average reading speed of the question wording excluding response options) have 4.2 higher odds of selecting top (positive) response options. This suggests that in the questions with vertical scale where response scale starts with positive options satisficing may lead to social desirability.

Selecting top options in a self-administered mode, primacy effect, has not been shown to be linked to social desirability previously, and has rarely been shown to be present in ordinal scales (the scale type we study) before. Our main interest is whether it is possible to increase the reporting of socially undesirable answers by reversing the scale such that the most negative option is at the top of the scale (see Figure 2).

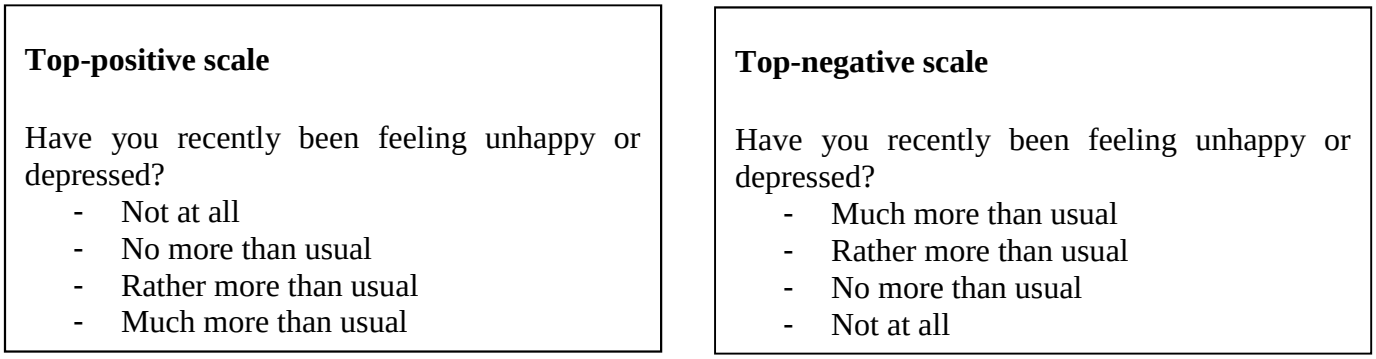

Figure 2. An example of a survey question in a control and treatment group

One of the standardized question sets that uses vertical response scale starting with a positive response option at the top is the 12-item General Health Questionnaire (GHQ-12). It is designed as a screening measure of minor psychological disorder (Goldberg, 1997), and is broadly used in a range of social surveys around the world, including in BHPS, mainstage UKHLS and in previous waves of IP. If satisficing leads to selecting top (positive) options in this scale, then respondents more prone to satisficing will appear to have better health than in reality. From previous research we know that satisficing is more likely among respondents

with less cognitive ability (Krosnick, 1991), which in turn is often related to lower education and income. This means that GHQ-12 as is implemented today may overestimate health for the more vulnerable groups.

In IP7 the GHQ-12 respondents were asked one of two versions as part of the self-completion section; either with positive response coming first, at the top of the scale, or with negative responses being presented first at the top. Households were randomly assigned to one of the two conditions. Finally, we included a second treatment providing a motivational message suggested by Cannel et al. (1981) with the following wording:

“In order for your answers to be most helpful to us, it is important that you try to be as thoughtful as you can. Since we need complete and accurate information from this research, we hope you will think hard to provide the information we need.”

GHQ-12 responses were recoded such that the most positive response always equalled one, and the most negative response always equal to four. Then the mean of the twelve questions were calculated for each respondent. This respondent mean is used to test the impact of this 2x2 experiment, and the means for each of the four combinations are presented in Table 27:

Table 27: Means of Respondent Means to GHQ-12 Scale by Conditions

	No Motivation	Motivation
Positive First	1.95	1.92
Negative First	1.93	1.92

These means shows little differences, and an ANOVA tests for mean differences (of respondent means) for each of the categorizations examining both experimental conditions and the interaction between the two confirm this lack of effect. For the main effect of the motivational statement, $F(1,2205) = 0.43$; $p=0.51$; for the main effect of the scale direction, $F(1,2205)= 0.64$, $p=0.42$; and for the interaction between the two, $F(1,2205) = 0.48$, $p=0.49$. Further analyses will examine each question more in-depth as well as subgroup analyses to identify possible differences.

j. Respondent Incentives (Peter Lynn)

At IP7, as at previous waves, sample members were sent an unconditional incentive with the advance letter notifying them of the upcoming wave of data collection. The incentives were in the form of a Love2Shop gift voucher. Within two parts of the sample, incentive treatments were allocated experimentally, while in a third part no experiment was carried out with incentives:

- In the mixed mode part of the sample the unconditional incentive was either £10 or £30. Also, some of those sent £10 in the mixed mode treatment group were promised an additional £20 for each household member conditional on all household members taking part online within two weeks of receiving the survey invitation. There were thus three treatment groups in this part of the sample;
- In the IP7 refreshment sample – all administered by CAPI – there were also three treatment groups: members received an unconditional incentive of £10, £20, or £30.
- In the CAPI part of the continuing sample, all sample members were provided a £10 incentive.

For all members of the continuing sample (original IP1 sample and the IP4 refreshment sample), the incentive provided at IP7 was the same incentive that they had received at IP6.

The level of incentive made a sizeable difference to response rate in the mixed mode sample, with individual response rates ranging from 59.3% with a £10 incentive to 66.2% with a £30 incentive amongst the IP4 refreshment sample. The range appears to have been even greater amongst the original (IP1) sample, where response rates were 57.1% with a £10 incentive and 67.3% with a £30 incentive (Figure 3). The data also suggest that the superior impact of the £30 unconditional incentive may have depended on the level of incentive that had been offered prior to IP6. The IP7 response rate was 63.4% amongst sample members whose incentive changed from £5 to £30 at IP6, but was 71.8% amongst those who incentive changed from £10 to £30 ($\chi^2(1)=3.54$; P = 0.06, results not shown).

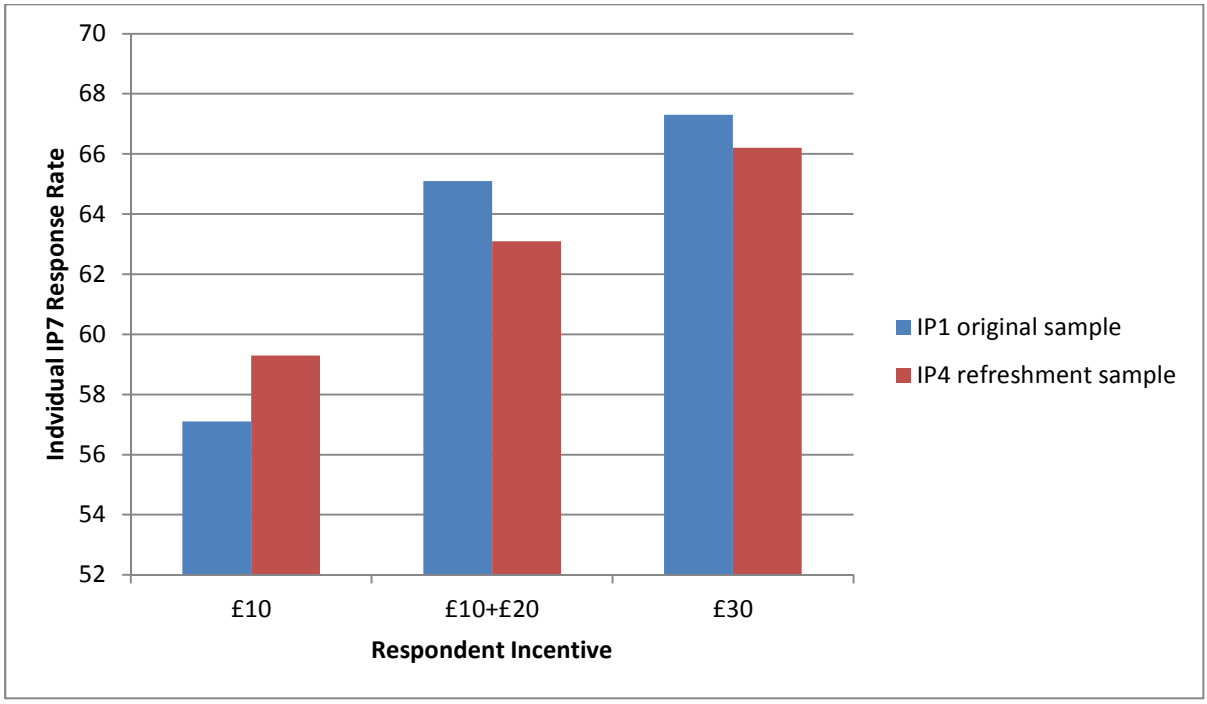

Figure 3: Individual response rates, by incentive treatment and sample

k. Digit Ratios in the IP7 Youth Sample (Sebastian Schnettler and Cara Booker)

Empirical research provides increasing evidence that hormones, social context, and prior behaviour interact in affecting behaviour throughout the life course in a number of behavioural domains such as violence and aggression, mating and parental investment, and status competition (Booth et al. 2006; Gettler et al. 2011; Harris 1999; Schnettler 2010; Schnettler and Nelson 2015; Taylor 2014). Yet, a scarcity of (representative) data sets that contain both hormonal and social context information hampers efforts to elucidate this complex interplay. Behavioural endocrinologists distinguish organizational hormonal effects through prenatal, in-utero hormone exposure from activational hormonal effects through serum concentrations of hormones. Organizational effects link early social context with developmental strategies and behavioural outcomes later in life by affecting brain development (Breedlove 2010; Hines 2011). Activational effects orchestrate behaviour in a more immediate way throughout the life course in response to changes of social context and prior behaviour.

Whereas serum concentrations of hormones can be directly assessed through blood and saliva samples and are increasingly used in behavioural experiments and observational studies, direct measurement of in-utero hormonal exposure is too risky. Thus, there is a need for indirect measurement. The length ratio of the index and ring finger (second and fourth digit, 2D:4D) has been found to be a stable marker of prenatal steroid hormone exposure with high validity and has been used extensively in small-scale behavioural studies (Breedlove 2010; Hönekopp et al. 2007; Manning et al. 1998). This marker has been employed in a large number of small-scale behavioural studies, providing indirect evidence for an association of prenatal hormonal exposure with various behavioural, health-related, personality, and physiological traits (Grimbos et al. 2010; Hönekopp, Voracek, and Manning 2006; Hönekopp and Watson 2011; Voracek et al. 2011; Voracek and Loibl 2009; Voracek, Pum, and Dressler 2010; Voracek, Tran, and Dressler 2010). Research also shows that important between- and within-gender differences exist in this association (Grimbos et al. 2010; Hönekopp and Watson 2010, 2011; Voracek et al. 2011).

Given a lack of large representative studies that combine digit ratio measurements with extensive background information on socioeconomic context and behavioural outcomes across the life course, we administered a 2D:4D module to a representative sample of 2,018 individuals in 1,187 household as part of wave 6 of the Understanding Society Innovation Panel (IP6). Hormonal information could thus be matched with longitudinal information on social context and behavioural outcomes in various life domains from multiple waves of the Innovation Panel. Deleting invalid cases (e.g. due to arthritis, broken fingers, and refusal to participate in the module) and outliers (values smaller and bigger than three standard deviations from the mean) left us with 1,468 right and 1,490 left hand measures of digit ratios for the adult sample. The digit ratio module was not administered to the youth sample in IP6. Instead, it was collected in Wave 7 of the Innovation Panel (IP7). Here we give a brief descriptive account of the resulting digit ratio measurements in IP 7. Whereas in IP6, digit ratios were measured by interviewers using callipers for the face-to-face sample and by respondents themselves in the web sample using whatever measurement device they had available at home, the adolescents in IP7 were asked to self-administer measurement of digit

lengths using a paper ruler attached to the self-completion questionnaire. Here we report on digit ratio measures in the IP7 youth sample.

For 31 of the 182 adolescents (86 boys, 96 girls) in the sample, either the left or right hand digit ratio measurement are missing (= 17.0%). For 29 respondents, both left and right hand measures are missing (= 16.0%). We further removed implausible values which we defined as finger lengths higher or lower than three standard deviations from the mean and digit ratios that are three standard deviations below or above the sex-specific mean. This leaves valid left hand measures for 147 respondents (= 80.8%) and valid right hand measures for 145 respondents (= 79.7%). The percentage of missing and implausible values is slightly higher for boys than for girls (22.1% vs. 16.7% for left and 22.1% vs. 18.8% for right hand measures). We also tested if missing values are more frequent, the younger the respondent child is. However, there is no clear association between birth year and the percentage of missing cases (see. Table 28).

Table 28: Percent missing and implausible cases in the digit ratio module, by year of birth

Birth year	***% missing cases***	
	(left hand)	***(right hand)***
1998	25.0	25.0
1999	10.5	5.3
2000	20.0	20.0
2001	36.0	44.0
2002	18.9	21.6
2003	19.2	19.2
2004	10.0	15.0

Right and left hand measures of the same individuals are only weakly correlated (Person's r = 0.164). Yet, this weak correlation is mainly due to the data of boys. Whereas in girls, left and right hand measures are moderately correlated with a correlation coefficient of about 0.465, the respective measures for boys are only weakly correlated with a correlation of about 0.118. A typical 2D:4D digit ratio for males is about 0.95 and for females about 1.0 (see Hönekopp and Watson 2010). We thus classify digit ratios for men and women as "masculine" (that is, indicating above-average testosterone exposure) if they are below the sex-specific average ratio and as "feminine" (that is, indicating below-average testosterone exposure) if they are

equal to or higher than the sex-specific ratio. We get a percentage of about 19.9% of discordant measurements, that is, cases in which left and right hand measurement leads to a different conclusion as to whether a digit ratio is to be classified as “male” or “female”.

Sex differences in mean digit ratios are not statistically significant. For the right hand, the mean digit ratio for girls is about 1.03 and that for boys about .994 ($t = 1.062$, $p = 0.291$). For the left hand, the respective values are 1.01 for girls and 1.05 for boys ($t = -0.966$, $p=0.337$). That is, in case of the right hand measures, the values are higher than those typically found in other studies, but the relative difference between girls and boys is as predicted: boys have more masculine digit ratios than girls on average, though these are not statistically significant. For left hand measures, however, mean values and the relative difference between the boys and girls are not as predicted in the literature (cf. Hönekopp and Watson 2010).

Overall, we thus conclude that measurement error in the youth sample may be too high in the digit ratio module of the IP7 youth sample to draw any valid conclusions. This is opposed to the digit module of the IP6 adult sample, where values were more closely aligned with previous research findings (cf. Al Baghal (ed). 2014).

References

Al Baghal, T. (ed.) (2014) Understanding Society Innovation Panel Wave 6: results from methodological experiments, Understanding Society Working Paper Series 2014-04. Colchester: University of Essex. https://www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2014-04

Andrews, F. M. (1984). Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach. *Public Opinion Quarterly*, *48*(2), 409–442.

Booth, A., Granger D.A, Mazur A., & Kivlighan, K.T. (2006). Testosterone and Social Behavior. *Social Forces* 85(1):167–91.

Breedlove, S. M. (2010). Minireview: Organizational Hypothesis: Instances of the Fingerpost. *Endocrinology* 151(9):4116–22.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. Sociological Methodology, 12, 389-437.

Castiglioni, L., Pforr, K., & Krieger, U. (2008). The Effect of Incentives on Response Rates and Panel Attrition: Results of a Controlled Experiment. *Survey Research Methods*, 2(3), 151-158.

Coates, J.M., Gurnell, M. & Rustichini, A. (2009). Second-to-Fourth Digit Ratio Predicts Success among High-Frequency Financial Traders. *Proceedings of the National Academy of Sciences* 106(2):623–28.

Couper, M.P. & Ofstedal, M.B. (2009). "Keeping in Contact with Mobile Sample Members" in P.Lynn (ed.) *Methodology of Longitudinal Surveys,* Chichester: John Wiley & Sons, Inc., 183-203.

Church, A. H. (1993). Estimating the effect of incentives on mail survey response rates: a meta-analysis. *Public Opinion Quarterly*, 57, 62-79Department of Health (2009). Shaping the Future of Care Together. London: The Stationery Office.

Gettler, L.T., McDade, T.W., Feranil, A.B, & Kuzawa, C.W. (2011). Longitudinal Evidence That Fatherhood Decreases Testosterone in Human Males. *Proceedings of the National Academy of Sciences* 108(39):16194–99.

Goldberg, D. P., Gater, R., Sartorius, N., Ustun, T. B., Piccinelli, M., Gureje, O., et al. (1997). The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychological Medicine, 27,* 191-197.

Goyder, J. (1987). *The Silent Minority: Non-respondents on sample surveys*. Oxford: Polity Press.

Grimbos, T, Dawood, K., Burriss, R.P., Zucker, K.J., & Puts, D.P. (2010). Sexual Orientation and the Second to Fourth Finger Length Ratio: A Meta-Analysis in Men and Women. *Behavioral Neuroscience* 124(2):278–87.

Harris, J.A. (1999). Review and Methodological Considerations in Research on Testosterone and Aggression. *Aggression and Violent Behavior* 4(3):273–91.

Hell, B. & Päßler, K. (2011). Are Occupational Interests Hormonally Influenced? The 2D:4D-Interest Nexus. *Personality and Individual Differences* 51:376–80.

Hönekopp, J., Bartholdt, L., Beier, L. & Liebert, A. (2007). Second to Fourth Digit Length Ratio (2D:4D) and Adult Sex Hormone Levels: New Data and a Meta-Analytic Review. *Psychoneuroendocrinology* 32(4):313–21.

Hönekopp, J., Voracek, M., & Manning, J.T. (2006). 2nd to 4th Digit Ratio (2D:4D) and Number of Sex Partners: Evidence for Effects of Prenatal Testosterone in Men. *Psychoneuroendocrinology* 31(1):30–37.

Hönekopp, J. & Watson, S. (2010). Meta-Analysis of Digit Ratio 2D:4D Shows Greater Sex Difference in the Right Hand. *American Journal of Human Biology* 22(5):619–30.

Hönekopp, J. & Watson, S. (2011). Meta-Analysis of the Relationship between Digit-Ratio 2D: 4D and Aggression. *Personality and Individual Differences* 51:381–86.

Hopkins, K. and Gullickson, A. (1992). Response Rates in Survey Research: A Meta-analysis of the Effects of Monetary Gratuities. *Journal of Experimental Education*, 61(1), 52-62.

Hui, C. H., & Triandis, H. C. (1985). The Instability of Response Sets. *Public Opinion Quarterly*, *49*(2), 253–260. http://doi.org/10.1086/268918

Jäckle, A. & Lynn, P. (2008). Respondent Incentives in a Multi-Mode Panel Survey: Cumulative Effects on Non-Response and Bias. *Survey Methodology*, 34(1), 105-117

James, W.H. (2001). Finger-Length Ratios, Sexual Orientation and Offspring Sex Ratios. *Journal of Theoretical Biology* 212(3):273–74

Kaminska, O. & Foulsham, T. (2013). "Understanding Sources of Social Desirability Bias in Different Modes: Evidence from Eye-Tracking." ISER working paper 2013-04; https://www.iser.essex.ac.uk/publications/working-papers/iser/2013-04

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

Laurie H, Smith R & Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics*, 15, 269-282.

Lippa, R.A. (2003). Are 2D: 4D Finger-Length Ratios Related to Sexual Orientation? Yes for Men, No for Women. *Journal of personality and social psychology* 85(1):179.

Lepkowski, J.M. & Couper, M.P. (2002). Non-response in the Second Wave of Longitudinal Household Surveys, in R. Groves et al. (eds), *Survey Non-Response*, New York: Wiley, 259-274,

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

Manning, J. T., Scutt, D., Wilson, J., & Lewis-Jones, D.I. (1998). The Ratio of 2nd to 4th Digit Length: A Predictor of Sperm Numbers and Concentrations of Testosterone, Luteinizing Hormone and Oestrogen. *Human Reproduction* 13(11):3000–3004.

McClendon, M. (1991). Acquiescence and Recency Response-Order Effects in Interview Surveys. *Sociological Methods & Research, 20*(1), 60–103.

McGonagle, K.. Couper, M.P. & Schoeni, R.F. (2011). Keeping Track of Panel Members: An Experimental Test of a Between-Wave Contact Strategy. *Journal of Official Statistics*, 27(2), pp. 319-338.

Putz, D.A., Gaulin, S.J.C, Sporter, R.J. & McBurney, D.H. (2004). Sex Hormones and Finger Length: What Does 2D:4D Indicate? *Evolution and Human Behavior* 25(3):182–99.

Schnettler, S. (2010). "Nature + Nurture = Love? A Test of the Trivers-Willard Hypothesis of Differential Parental Investment on the Basis of Sociological and Biological Explanations." Dissertation, Yale University, Ann Arbor, MI: UMI.

Schnettler, S. & Nelson, S. (2015). Aggression Und Gewalt. in B. P. Lange and S. Schwarz (eds). *Die menschliche Psyche zwischen Natur und Kultur*, p. 150–59 Lengerich: Papst Publishers.

Schuman, H., & Ludwig, J. (1983). The norm of even-handedness in surveys as in life. *American Sociological Review*, *48*(1), 112-20.

Schuman, H., & Presser, P. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Shanahan, M.J., Hofer, S.M. & Shanahan, L. (2003). Biological Models of Behavior and the Life Course. in J. T. Mortimer and M. J. Shanahan (eds). *Handbook of the life course*. New York: Kluwer Academic Publishers.

Singer, E., Van Hoewyk, J., & Gebler, N. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics*, 15, 217-230.

Taylor, C.J. (2014). Physiological Stress Response to Loss of Social Influence and Threats to Masculinity. *Social Science & Medicine* 103:51–59.

Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge, United Kingdom: Cambridge University Press

Voracek, M and Mariella Loibl, L. (2009). Scientometric Analysis and Bibliography of Digit Ratio (2D:4D) Research, 1998-2008. *Psychological Reports* 104(3):922–56.

Voracek, M., Pietschnig, J., Nader, I.W., & Stieger, S. (2011). Digit Ratio (2D:4D) and Sex-Role Orientation: Further Evidence and Meta-Analysis. *Personality and Individual Differences* 51(4):417–22.

Voracek, M., Pum U., & Dressler, S.G. (2010). Investigating Digit Ratio (2D:4D) in a Highly Male-Dominated Occupation: The Case of Firefighters. *Scandinavian Journal of Psychology* 51(2):146–56.

Voracek, M., Tran, U.S, & Dressler, S.G. (2010). Digit Ratio (2D:4D) and Sensation Seeking: New Data and Meta-Analysis. *Personality and Individual Differences* 48(1):72–77.

Wanless, D. (2006). Securing good care for older people. Taking a long-term view. London: The King's Fund.

Appendix

For the 6 items, "positive" and "negative" formulations. "positive" means that with a disagree-agree scale, the socially desirable direction will be towards the higher end of the scale.

Trait number	SD direction	Item formulation
T1	Positive	The UK should allow **more** people of the same race or ethnic group as most British people to come and live here
T2	Positive	UK should allow **more** people of a different race or ethnic group from most British people to come and live here
T3	Positive	UK should allow **more** people from the poorer countries outside Europe to come and live here
T4	Positive	It is generally **good** for UK's economy that people come to live here from other countries
T5	Positive	UK's cultural life is generally **enriched** by people coming to live here from other countries
T6	Positive	UK is made a **better** place to live by people coming to live here from other countries
T1	Negative	The UK should allow **fewer** people of the same race or ethnic group as most British people to come and live here
T2	Negative	UK should allow **fewer** people of a different race or ethnic group from most British people to come and live here
T3	Negative	UK should allow **fewer** people from the poorer countries outside Europe to come and live here
T4	Negative	It is generally **bad** for UK's economy that people come to live here from other countries
T5	Negative	UK's cultural life is generally **undermined** by people coming to live here from other countries
T6	Negative	UK is made a **worse** place to live by people coming to live here from other countries