Understanding Society

Working Paper Series

No. 2016-02

July 2016

Understanding Society Innovation Panel Wave 8: Results from Methodological Experiments

Tarek Al Baghal (ed.)

Contributors: Mathew Creighton[3], Jennifer Dykema[5], Alessandra Gaia[1], Alexandru Cernat[4], Dana Garbarski[7], Amaney Jamal[8], Olena Kaminska[1], Florian Keusch[2], Peter Lynn[1], Daniel Oberski[6], Nora Cate Schaeffer[5], S.C. Noah Uhrig[9], Ting Yan[2]

[1] Institute of Social and Economic Research, University of Essex, [2] University of Michigan, [3] University of Massachusetts-Boston, [4] University of Manchester [5] University of Wisconsin, [6] Tilburg University, [7] Loyola University-Chicago, [8] Princeton University, [9] Ministry of Justice (UK)

Understanding Society Innovation Panel Wave 8: Results from Methodological Experiments

Tarek Al Baghal (ed.)
Contributors: Mathew Creighton[3], Jennifer Dykema[5], Alessandra Gaia[1], Alexandru Cernat[4], Dana Garbarski[7], Amaney Jamal[8], Olena Kaminska[1], Florian Keusch[2], Peter Lynn[1], Daniel Oberski[6], Nora Cate Schaeffer[5], S.C. Noah Uhrig[9], Ting Yan[2]

[1] Institute of Social and Economic Research, University of Essex, [2] University of Michigan, [3] University of Massachusetts-Boston, [4] University of Manchester [5] University of Wisconsin, [6] Tilburg University, [7] Loyola University-Chicago, [8] Princeton University, [9] Ministry of Justice (UK)

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 8 of the Innovation Panel in the spring and summer of 2015.

IP8 was the fourth 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 through IP7, IP8 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 have been surveyed via face-to-face interviews only. 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.

As with prior waves, there was a methodological experiment involving the amount of respondent incentives. Further experiments examine the measurement of attitudes on sensitive issues using a technique using item counts, interviewers' assessment of respondents' health, taking multiple measurements to better understand attitudes and the impact of how scales are presented.

Understanding Society Innovation Panel Wave 8: Results from Methodological Experiments

Tarek Al Baghal (ed.)

Contributors: Mathew Creighton[3], Jennifer Dykema[5], Alessandra Gaia[1], Alexandru Cernat[4], Dana Garbarski[7], Amaney Jamal[8], Olena Kaminska[1], Florian Keusch[2], Peter Lynn[1], Daniel Oberski[6], Nora Cate Schaeffer[5], S.C. Noah Uhrig[9], Ting Yan[2]

[1] Institute of Social and Economic Research, University of Essex, [2] University of Michigan, [3] University of Massachusetts-Boston, [4] University of Manchester [5] University of Wisconsin, [6] Tilburg University, [7] Loyola University-Chicago, [8] Princeton University, [9] Ministry of Justice (UK)

Abstract

This paper presents some preliminary findings from Wave 8 of the Innovation Panel (IP8) of *Understanding Society*: The UK Household Longitudinal Study. *Understanding Society* is a major panel survey in the UK. May 2015, the eighth wave of the Innovation Panel went into the field. IP8 used a mixed-mode design, using on-line interviews and face-to-face interviews. This paper describes the design of IP8, 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 eighth wave of the Innovation Panel (IP8) of *Understanding Society*: The UK Household Longitudinal Study (UKHLS). *Understanding Society* is a major panel survey for the UK. The first six waves of data collection on the main sample have been completed, and seventh and eighth waves are currently in the field. The data from the first five waves of the main samples are available from the UK Data Archive, and the sixth 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 seven 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. The IP7 started fieldwork in June 2015 and added 488 responding households as a refreshment sample. Working Papers which cover the experimentation carried out in all seven innovation panels are available from the *Understanding Society* website.[2] The data from the first seven waves of the Innovation Panel are held at the UK Data Service. This paper describes the design of IP8, 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 IP8. Section 4 then reports on the experiments carried at IP8.

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 Household Panel Survey (BHPS) and other national panels around the world. In the first

[2] https://www.understandingsociety.ac.uk/research/publications/working-papers

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 laptops 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). For the first seven waves, surveys of continuing sample members were interviewer-administered. At Wave 8 it was decided that the 20% of household identified as having the lowest likelihood of responding in the mixed-mode would be assigned immediately to the CAPI-only design, while the remaining 80% would be randomly allocated to one the two designs. The end result is about 60% of households will be CAPI-only and 40% will be mixed-mode.

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. Innovation Panel Wave 8: Design

IP8 was comprised of three samples: the original sample from IP1, the IP4 refreshment sample, and the IP7 refreshment sample. IP8 employed a mixed-mode design, which started in IP5 has been used in each subsequent wave. 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 has been maintained at each wave. However, at IP8 subgroup of households with a very low propensity to respond via the web in in the CAWI condition were assigned to CAPI to begin fieldwork. Very low web propensity was determined by modelling web-completion using IP5, IP6, and IP7 data. The IP7 refreshment sample units were all allocated to the F2F design. TNS BMRB conducted fieldwork at IP7 and IP8, after the first six waves were conducted by NatCen.

There was a "soft" launch of the CAWI phase, consisting of 100 of the CAWI-first households to identify any problems, with the "main" launch consisting of the remaining households occurring one week later. 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. A reminder letter was then sent to all adults in the WEB group who had not completed their interview. This letter was sent two weeks after the initial advance letter for the main CAWI launch.

After nearly three weeks of the main CAWI launch being in the field, all adults who had not

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

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 weeks (eight for the soft launch) the CAWI survey was closed, and only CAPI surveys were conducted until the mop-up phase. The main CAPI fieldwork lasted 16 weeks, after which the mop-up phase started.

The WEB-only period ran from 6th May to 1st June for the soft launch households and 12th May to 1st June for the main CAWI sample households. The face-to-face fieldwork started 2nd June and ran until 16th September. Interviewers could continue to attempt CAPI surveys during the mop-up period. The mop-up follow-up phase attempted interviews with those not responding in both the WEB and F2F versions, through CAPI, CATI or CAWI available. This final phase ran from 17th September 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 across Great Britain. The briefings took place between 13th May and 28th May 2015, with a total of 120 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

IP8 was the sixth 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 17th March with a deadline of 11th April. Twenty-two proposals were received with four being accepted. There were for a total of eight experiments included in IP8; in addition the four new experiments, three were carried over from IP7 and one from IP6 (on mode preference).. The twenty-two 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.

b. Sample

There were three sample issued at IP8: the original sample; the refreshment sample from IP4; and the refreshment sample issued at IP7. These samples were comprised of those households who had responded at IP7, plus some households which had not responded at IP7. 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 840 original sample households, 399 IP4 refreshment sample households and 499 IP7 refreshment sample households issued at IP7. There were 1738 total sample households issued at IP8. All of the households were originally selected from the Postcode Address File (PAF) using the same methods.[4]

As discussed above, around two-thirds of the original and IP4 refreshment samples were allocated to the mixed-mode design in IP5, which has been maintained all subsequent waves, including IP8. 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. As noted, some households that were deemed to have a very low propensity to respond via were assigned directly to CAPI to begin fieldwork. 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.

[4] See Lynn, P. (2009). Sample Design for Understanding Society *Understanding Society Working Paper Series No. 2009 – 01* at https://www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2009-01

Table 1: Allocation to mode design by sample type

	Original Sample	IP4 Refreshment Sample	Total
CAPI only	322 38.3%	160 40.1%	482 38.9%
Mixed-mode (CAWI+CAPI)	518 61.7%	239 59.9%	757 61.1%
Total	840	399	1,239

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.[5] The interview included the following 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, 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.

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.[6] 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

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

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 IP8 as a whole. Section g describes the effect of incentives on response rates. The issued sample at the eighth wave consisted of 1582 households that had responded to IP7 and 156 households that had not responded at IP7. For the original and IP4 refreshment samples, IP7 continued with the mixed-mode design experiment implemented since IP5, and the same sample allocation (CAPI-only or a Web and CAPI mixed-mode design) was maintained, with the noted households estimated to have low propensity to respond by web being assigned to CATI. Fieldwork for the IP7 refreshment sample used a CAPI-only design for both IP7 and IP8.

Table 2 displays the household-level response at IP8 for the original and IP4 refreshment samples by CAPI-only and mixed-mode conditions and the overall total response. The lower panel displays individual response rate for each. For each cell, the percent is reported above the number of units the percent represents, in italics. The total number of eligible sampled units is in the Total rows, in bold.

Table 2. Household and Individual Response Outcomes for Original and IP4 Refreshment samples, IP8

	Original Sample		IP4 Refreshment Sample		Combined		Total
Household RR	F2F	MM	F2F	MM	F2F	MM	
Complete HH	60.8% *186*	67.9% *336*	58.7% *91*	66.7% *150*	60.2% *278*	67.5% *486*	64.6% *763*
Partial HH	16.7% *51*	18.6% *92*	16.1% *25*	20.0% *45*	16.5% *76*	19.0% *137*	18.0% *213*
Total Responding HH	77.5% *237*	86.5% *428*	74.8% *116*	86.7% *195*	76.6% *354*	86.5% *623*	82.7% *976*
Nonresponding HH	22.5% *69*	13.5% *67*	25.2% *39*	13.3% *30*	23.4% *108*	13.5% *97*	17.3% *205*
Total HH	**307**	**495**	**155**	**225**	**462**	**720**	**1181**
Conditional Individual RR	F2F	MM	F2F	MM	F2F	MM	
Responding individuals	84.8% *378*	85.8% *751*	86.0% *222*	84.7% *342*	85.2% *569*	85.5% *1093*	85.4% *1662*
Nonresponding individuals	15.3% *68*	14.2% *124*	14.0% *31*	15.4% *62*	14.8% *99*	14.5% *186*	14.6% *285*
Total Ind.	**446**	**875**	**222**	**404**	**668**	**1279**	**1947**

There were 976 interviewed households from the continuing samples, for an 82.7% overall household response rate. Within these households, 1662 people were interviewed, for a conditional individual response rate of 85.4%.

Table 3 shows the household-level and individual-level response at IP8 for the IP7 refreshment sample, conducted only by face-to-face interviews. For the IP7 refreshment sample, 374 households were surveyed, a 76.1% response rate. Of all of the enumerated individuals in these households, 605 were interviewed, equalling an 83.8% response rate.

Table 3. Household and Individual Response Outcomes for IP7 Refreshment sample, IP8

	IP7 Refreshment Sample (Initial wave – CAPI only)
Household RR	
Complete HH	58.7% *288*
Partial HH	17.5% *86*
Total Responding HH	76.1% *374*
Nonresponding HH	23.8% *117*
Total HH	**491**
Conditional Individual RR	
Responding individuals	83.8% *605*
Nonresponding individuals	16.2% *117*
Total Individuals	722

The response rates for IP7 refreshment sample are similar to other samples at IP8 assigned to the face-to-face only condition. The household response rates, based on households that had responded at some prior wave, are consistently higher for those assigned to the mixed-mode conditions relative to any of the F2F samples. However, this is not the case for the individual conditional response rates. Once a household has accepted the survey request for either mode condition, the individuals within these households respond at similar rates.

Given the mixed-mode design used for portions of the original and IP4 refreshment samples at IP8, not all individuals responded in the same mode. Further, at IP8 the mop-up period was again used, where non-responding units all the samples were contacted and could respond via the web or by telephone, regardless of the allocated mode design. Table 4 shows the mode of completion for individuals in these three samples by mixed-mode condition (for IP1 and IP4 samples) and total overall at IP8 including the mop-up phase. Given the similarities in response rates the IP7 refreshment sample were included with the original and IP4 refreshment samples in the combined figures.

Table 4. Mode of Response, IP8

	Original Sample		IP4 Refreshment Sample		IP7 Refreshment Sample	Combined		Total
Responding Mode	F2F	MM	F2F	MM	F2F	F2F	MM	
Face-to-Face	95.5% *361*	31.2% *234*	93.2% *178*	25.2% *86*	95.9% *580*	95.3% 1119	29.3% *320*	63.5% *1439*
Web	3.7% *14*	67.2% *505*	4.2% *8*	74.0% *253*	3.1% *19*	3.5% *41*	69.4% *758*	35.2% *799*
Telephone	0.8% *3*	1.6% *12*	2.6% *5*	0.9% *3*	1.0% *6*	1.2% *14*	1.4% *15*	1.3% *29*
Total Ind.	**378**	**751**	**191**	**342**	**605**	**1174**	**1093**	**2267**

IP8 was the first wave where it was possible to access the web survey using any internet-enabled device. In previous waves, smartphones were blocked from accessing the survey, although tablets could access the questionnaire. A number of variables were captured about the device the survey was accessed with, including what type of device was used, the operating system, the device model, the browser used, browser version, and screen resolution. These variables are now available in IP7 and IP8 as w_deviceused w_deviceos w_devicemodel w_browserused w_browserversion w_screenresolution in the file w_indresp_ip. The distribution of devices used across all samples in IP8 is presented in Table 5.

Table 5. Device Used, Web Respondents, Wave 8

	IP8 Web Respondents
PC/Laptop	70.4% *562*
Large Tablet	19.5% *156*
Small/Medium Tablet	3.5% *28*
Smartphone	4.1% *33*
Other	2.5% *20*
Total Web Respondents	**799**

Longitudinal Response Outcomes

The individual re-interview rate is an important outcome in a longitudinal survey, since analyses require pairs of observations to measure change. Re-interview rates are calculated as the percentage of eligible units responding at later waves who were also surveyed at the initial wave. For those in the original sample, the percentage is predicated on response at IP1, while the fourth wave is the initial wave for the IP4 refreshment sample, and the seventh wave was the first for IP7.

Table 6 presents the longitudinal individual re-interview rates for the original sample (for IP2-IP7), the IP4 refreshment sample (for IP5-IP7), and IP7 (for IP8). For each cell, the percent is reported above the number of individuals the percent represents, in italics.

Table 6. Longitudinal re-interview rates

	IP2	IP3	IP4	IP5	IP6	IP7	IP8
Original Sample	69.3% *1654*	60.6% *1442*	54.7% *1270*	45.9% *1095*	45.9% *1100*	38.4% *917*	36.2% *867*
IP4 Refreshment Sample	-	-	-	82.0% *586*	76.8% *554*	62.1% *447*	58.8% 423
IP7 Refreshment Sample							79.2% *520*

As with any longitudinal study, there has been attrition at each wave, decreasing the overall numbers for each sample. At IP8, 867 individuals from the original sample who responded at IP1 were successfully interviewed, representing a 36.2% re-interview rate. For the IP4 refreshment sample, the IP7 was their fourth wave and 423 responded, for a 58.8% re-interview rate. IP8 was the second wave for the IP7 refreshment sample, with 520 responses for a 79.2% re-interview rate.

4. Experimentation in IP8

There were a number of experiments carried on IP8 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 IP8; 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. Masking opposition to immigration: an experimental approach to understand the dynamics of social desirability bias (Mathew J. Creighton, Amaney Jamal)

There is a growing body of literature showing that intolerance is masked from direct questioning. We use the Item Count Technique (ICT), also known as the list experiment, to manipulate the level of anonymity offered respondents. As a comparison, we pose a direct question to an additional group that corresponds to the questions measured with the list experiment. First, using the direct questions, we estimate the proportion of the population in the UK who openly express opposition to three distinct types of immigrants, defined by characteristics of the country of origin – Muslim, Eastern European and Caribbean. This defines the overt opposition these immigrant groups confront. Second, using the ICT, we estimate the proportion who anonymously expresses opposition to immigrants from the same country-type origins. This defined the covert opposition these same groups confront. Third, we compare the overt and covert proportions to ascertain the proportion of the population that masks their opposition. This captures the level of social desirability bias (SDB).

The Measures:

The Direct Questions:

The following three direct questions are posed to an independent sample of respondents.

Direct 1:

Do you think the UK should allow people from Muslim countries to come and live here?

- Allow to come and live here
- Do not allow to come and live here

Direct 2:

Do you think the UK should allow people from Eastern European countries to come and live here?

- Allow to come and live here
- Do not allow to come and live here

Direct 3:

Do you think the UK should allow people from Caribbean countries to come and live here?

- Allow to come and live here
- Do not allow to come and live here

The ICT:

The following question was posed to an independent sample of respondents, referred to as the control group.

Control List:

Of the following three statements, HOW MANY of them do you AGREE with? We don't want to know which statements, just HOW MANY.

- The UK should increase assistance to the poor
- The UK should decrease the tax on diesel and petrol
- The UK should allow large corporations to pollute the environment

The following three questions were posed to three independent samples each of which constitute a treatment group.

Treatment List 1:

Of the following four statements, HOW MANY of them do you AGREE with? We don't want to know which statements, just HOW MANY.

- The UK should increase assistance to the poor
- The UK should decrease the tax on diesel and petrol
- The UK should allow large corporations to pollute the environment

- The UK should allow people from Muslim countries to come and live here

Treatment List 2:

Of the following four statements, HOW MANY of them do you AGREE with? We don't want to know which statements, just HOW MANY.

- The UK should increase assistance to the poor
- The UK should decrease the tax on diesel and petrol
- The UK should allow large corporations to pollute the environment
- The UK should allow people from Eastern European countries to come and live here

Treatment List 3:

Of the following four statements, HOW MANY of them do you AGREE with? We don't want to know which statements, just HOW MANY.

- The UK should increase assistance to the poor
- The UK should decrease the tax on diesel and petrol
- The UK should allow large corporations to pollute the environment
- The UK should allow people from Caribbean countries to come and live here

The Method:

The preliminary analysis consists of three steps. The first estimates the overt opposition. This is straightforward as the question is directly posed to an independent sample of respondents and can be derived directly from the response to the question (Direct 1, Direct 2 and Direct 3 above). We'll refer to this as$\bar{X}_D$. The second step derives the covert opposition by subtracting the mean response pattern to each of the three list questions (Treatment List 1, Treatment List 2 and Treatment List 3 above) from the mean response to the control list question (Control List above) using equation (1):

$$Z = \bar{X}_L - \bar{X}_C \qquad (1)$$

where Z is the proportion of the sample that select the additional list item in the treatment, which is derived from the difference between the mean response to the treatment, defined by the indicator L, and the mean response to the control list, defined by the indicator C.

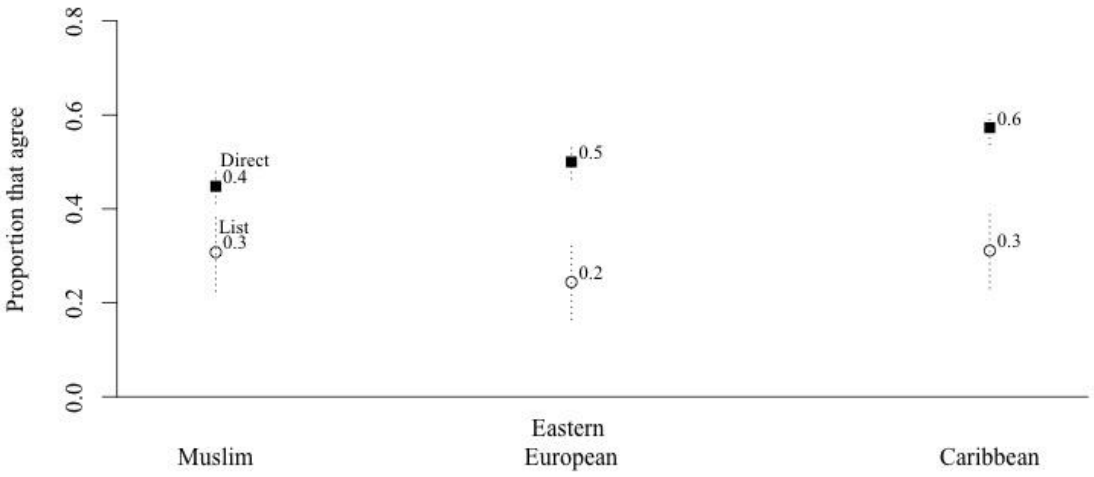
The third step is the estimation of the extent to which opposition is masked. This is done using the ICT, expressed by equation (2):

$$B = \bar{X}_D - Z \qquad (2)$$

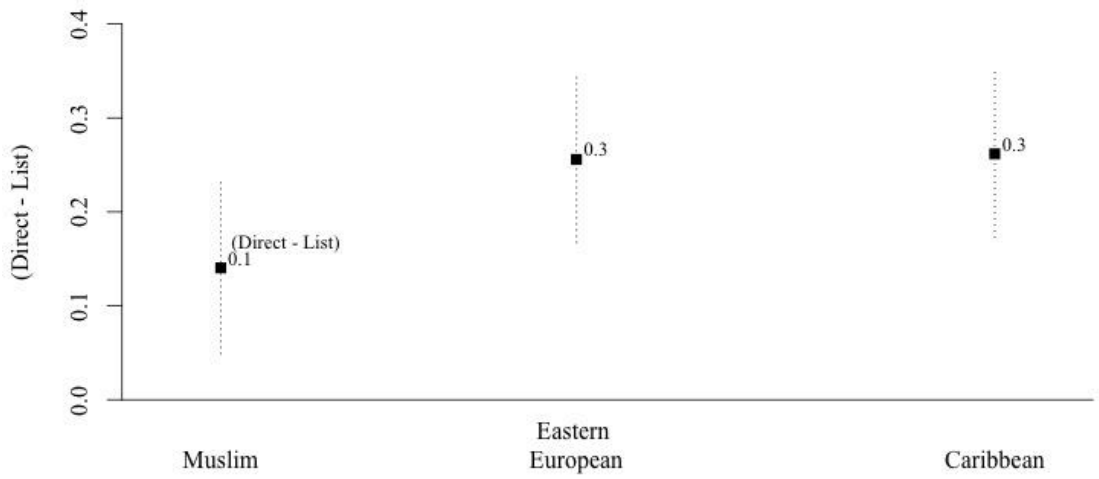
where B is direct measure of SDB that, when converted to a percentage scale, is typically interpreted as the number of percentage points difference between the explicit, derived from the control sample, and the implicit estimate (Z), derived from equation (1).

Preliminary Results:

Plot 1: Muslim, Eastern European and Caribbean immigrants should be allowed to come and live in the UK



Plot 2: Social Desirability Bias



Plot 1 reports the estimated proportion in favor of allowing immigrants of each of the three country-origin types assessed by the experiment. The direct measure ($\bar{X}_D$), which is higher in each case, is directly estimated. The list proportion (Z) is estimated using equation (1). Plot 2 shows the levels of Social desirability bias for each of the three country-origin types assessed in the experiment. SDB (B) is derived using equation (2).

b. A comparison of self-reported sexual identity using direct and indirect questioning (Alessandra Gaia, Noah Uhrig)

This experiment aims at addressing the following research questions: What is the estimated prevalence of the lesbian gay and bisexual population obtained with an indirect questioning method, such as the "Item Count" indirect questioning Technique (ICT)?

Does protocol involving face-to-face interviewing with a show card lead to underreporting of sexual minority status compared to a computer administered self-interview (CASI) protocol? How do these two estimates compare with the estimate produced using the "Item Count" Technique (ICT)?

Does the indirect questioning technique reduce the ratio of non-usable to usable data when estimating sexual minority population sizes compared to either interviewer administered or self-administered direct questioning?

Method

Using a Two-List Item Count Technique (ICT), we measure sexual attraction and sexual identity. In the "Two-List" ICT, respondents are randomly assigned to one of the two groups, but every individual receives two lists. For one group the sensitive item is included in the first list but not the second, for the other group the sensitive item is included in the second list but not the first (Tourangeau *et al.* 2001). "The difference in the mean number of items reported by the two groups is the estimated proportion" of the sensitive characteristic (Tourangeau and Yan 2007:872).

The Two-List ICT is crossed with random allocation into groups receiving versions of either the UKHLS or the Integrated Household Study (IHS) direct sexual identity question. The UKHLS adopts a self-completion approach whereas the IHS uses an interviewer administered approach with a show card, if face-to-face, or no show card when over the telephone[7].

We separated the ICT list questions from the direct sexual identity question in the questionnaire in order to avoid carry-over effects between these survey tasks. The IP8 mixed-

---

[7] More specifically, the UKHLS question adopts the following wording: "Which of the following options best describes how you think of yourself?" "Heterosexual or Straight", "Gay or Lesbian", "Bisexual", "Other", and "Prefer not to say". In addition to these categories, respondents could refuse to answer ("Don't want to answer") or report "don't know"; however, these two options became visible only once the respondent attempted to skip the survey question leaving the fields empty. Given the self completion nature of this question, this is not asked in telephone interviews. We refer to this protocol as "UKHLS", as this is the protocol currently adopted (among others) by the Understanding Society UK Household Longitudinal Study.

The IHS question is worded as following: "Which of the options on this card best describes how you think of yourself? Please just read out the number next to the description." Response categories are: "Heterosexual or Straight", "Gay or Lesbian" "Bisexual", and "Other". In addition to these categories, respondents could refuse to answer ("Don't want to answer") or report "doesn't know".

This protocol is also asked in telephone interview, as following: "I will now read out a list of terms people sometimes use to describe how they think of themselves: "Heterosexual or Straight", "Gay or Lesbian", "Bisexual", or "Other". As I read the list again please say 'yes' when you hear the option that best describes how you think of yourself." We refer here to this protocol as "IHS", as this is the protocol adopted in the Integrated Household Survey.

mode design was independent of this experiment, though mode allocation was catered for in the allocation to direct questioning design protocol.

Results

Overall, sample members reacted well to the ICT questions on sexual orientation. On all questions item non response was low, with only one respondent skipping the survey question. Refusal was also not frequent, ranging from 3.3% (n=37) of respondents to 0.55% (n=6) respondents; and don't know answers were rare, ranging from to 0.65% (n=7) to 0.19% (n=2).

The questions were designed so that the list of items would fit together and make sense to the respondent – as suggested by Droitcour et al. (1991). Moreover, the lists were designed to have a mix of "low prevalence" and "high prevalence" items. Indeed, if all items in the list are of a high prevalence, the respondent may count all items in the list, and thus self-identify ("ceiling effect"); conversely, if all "non-sensitive" items are very rare, the respondent may fear that by counting one item, he would similarly self-identify ("floor effect").

Thus, we combined items that we expected to be low prevalence (e.g. "I would describe myself as being disabled"), with items that we expected to be high prevalence (e.g. "I would describe myself as being British").

Unexpectedly, in the fields of attraction (lists A and B) and behaviour (lists C and D), the relative majority (over 36%) of respondents reported that none of the items presented applied to them; thus, we have evidence of a "floor effect"; conversely, in the identity questions (lists E and F) the "floor effect" was not problematic, as "none of the statements are true" was selected by only a tiny percentage of respondents.

The evidence on the "ceiling effect" is mixed; while lists A (attraction), C (behaviour) and E (identity) appear well designed, with only a tiny proportion of respondents selecting that all "four statements are true"; conversely, in lists B (attraction), list D (behaviour) and F (identity), the prevalence of respondents reporting all four behaviours ranges between 7 and 20%, indicating that a non-ignorable fraction of respondents may have not revealed the sensitive item in the full list (the one including the sensitive item) to avoid disclosing the sensitive attribute.

Both “ceiling” and “floor” effects may have influenced the estimates of the “attraction” and “identity” items, where, unexpectedly, we observed a lower average in the list with the sensitive item (“List B+S”, “List E+S”, “List F+S”), compared with the average in the list without the sensitive items – “List B”, “List E”, and “List F” (see table 8).

Vice versa, and consistently with our expectations, in the “behavioural” questions we observed a higher average in the lists which include the sensitive item (“List C+S” and “List D+S”), compared with the list that excludes the sensitive item (“List C” and “List D”). The resulting estimated prevalence of the population having had a homosexual sexual experience is 9.9% (see table 8).

Table 7. Item Count Technique: descriptive statistics

ATTRACTION		LIST A		LIST A+S	
I have at least once been sexually **attracted** to someone who …		Obs	%	Obs	%
Missing		*1*	*0.09*		
Refusal		*32*	*2.97*	*10*	*0.92*
don't know		*3*	*0.28*	*4*	*0.37*
• is the same sex as me	None of the statements are true	401	37.2	392	36.23
• has a disability	One of the statements are true	274	25.42	265	24.49
• is fit and muscular	Two of the statements are true	206	19.11	231	21.35
• grew up with me in my local area	Three of the statements are true	129	11.97	120	11.09
• is ten or more years older than me	Four of the statements are true	32	2.97	45	4.16
How many statements are true for you?	Five statements are true	N.A.	N.A.	15	1.39

		LIST B		LIST B+S	
I have at least once been sexually **attracted** to someone who …		Obs	%	Obs	%
missing				*1*	*0.09*
refusal		*10*	*0.92*	*29*	*2.69*
don't know		*6*	*0.55*	*7*	*0.65*
• is the same sex as me	None of the statements are true	419	38.72	450	41.71
• wears the latest trends and fashions	One of the statements are true	204	18.85	192	17.79
• has a tattoo or body piercing	Two of the statements are true	190	17.56	144	13.35
• is of a different ethnicity to me	Three of the statements are true	113	10.44	115	10.66
• is from a different class background to me	Four of the statements are true	140	12.94	111	10.29
How many statements are true for you?	Five statements are true	N.A.	N.A.	30	2.78

Table 7. Item Count Technique: descriptive statistics (continued)

BEHAVIOUR		LIST C		LIST C+S	
I have at least once had an **experience** of a sexual kind – for example kissing, cuddling or sexual intercourse – with a person who …		Obs	%	Obs	%
missing		*1*	*0.09*		
refusal		*37*	*3.43*	*17*	*1.57*
don't know		*2*	*0.19*	*5*	*0.46*
• is the same sex as me	None of the statements are true	453	42.02	448	41.4
• has a disability	One of the statements are true	304	28.2	295	27.26
• is fit and muscular	Two of the statements are true	196	18.18	191	17.65
• grew up with me in my local area	Three of the statements are true	73	6.77	84	7.76
• is ten or more years older than me	Four of the statements are true	12	1.11	35	3.23
How many statements are true for you?	Five statements are true	N.A.	N.A.	7	0.65

		LIST D		LIST D+S	
I have at least once had an **experience** of a sexual kind – for example kissing, cuddling or sexual intercourse – with a person who …		Obs	%	Obs	%
missing				*1*	*0.09*
refusal		*15*	*1.39*	*36*	*3.34*
don't know		*4*	*0.37*	*5*	*0.46*
• is the same sex as me	None of the statements are true	517	47.78	516	47.87
• wears the latest trends and fashions	One of the statements are true	233	21.53	182	16.88
• has a tattoo or body piercing	Two of the statements are true	154	14.23	147	13.64
• is of a different ethnicity to me	Three of the statements are true	79	7.3	106	9.83
• is from a different class background to me	Four of the statements are true	80	7.39	62	5.75
How many statements are true for you?	Five statements are true	N.A.	N.A.	23	2.13

IDENTITY		LIST E		LIST E+S	
I would describe myself as being …		Obs	%	Obs	%
missing		*1*	*0.09*		
refusal		*8*	*0.74*	*10*	*0.92*
don't know		*3*	*0.28*	*4*	*0.37*
• gay, lesbian or bisexual	None of the statements are true	40	3.71	50	4.62
• stylish and fashionable	One of the statements are true	237	21.99	276	25.51
• disabled	Two of the statements are true	538	49.91	485	44.82
• patient	Three of the statements are true	235	21.8	242	22.37
• British	Four of the statements are true	16	1.48	14	1.29

How many statements are true for you?	Five statements are true	N.A.	N.A.	1	0.09
		LIST F		LIST F+S	
I would describe myself as being …		Obs	%	Obs	%
missing				*1*	*0.09*
refusal		*6*	*0.55*	*12*	*1.11*
don't know		*3*	*0.28*	*3*	*0.28*
• gay, lesbian or bisexual	None of the statements are true	28	2.59	46	4.27
• healthy	One of the statements are true	163	15.06	173	16.05
• tolerant	Two of the statements are true	279	25.79	296	27.46
• European	Three of the statements are true	384	35.49	384	35.62
• working class	Four of the statements are true	219	20.24	157	14.56
How many statements are true for you?	Five statements are true	N.A.	N.A.	6	0.56

Table 8. The estimates from the Item Count Technique

Average "List A"	Average "List A+S"	Average "List A+S" – Average "List A"
1.153	1.257	0.104
Average "List B"	Average "List B+S"	Average "List B+S" – Average "List B"
1.391	1.362	-0.029
Estimated prevalence of homosexual/bisexual attraction: N.A.		
Average "List C"	Average "List C+S"	Average "List C" – Average "List C+S"
0.928	1.042	0.114
Average "List D"	Average "List D+S"	Average "List D" – Average "List D+S"
1.033	1.117	0.084
Estimated prevalence of homosexual/bisexual experience: 9.9%		
Average "List E"	Average "List E+S"	Average "List E+S" – Average "List E"
1.953	1.904	-0.0502
Average "List F"	Average "List F+S"	Average "List F+S" – Average "List F"
2.562	2.425	-0.137
Estimated prevalence of homosexual/bisexual identity: N.A.		

In addition to the Item Count Technique experiment, we also compare two protocols for asking sexual identity: the self completion "UKHLS" protocol and the face to face with showcard "IHS" protocol. As shown in table 9 there are no statistically significant difference across the two protocols.

Also, comparing the two protocols for males and females separately, and within different age groups (16-24, 25-34, 35-44, 45-54, 55-64 and 65+) we observed that the two protocols do not result in different estimates (results not shown).

Table 9. A comparison of the UKHLS and IHS protocols

	UKHLS				IHS			
	Obs.	%	95% C.I.		Obs.	%	95% C.I.	
heterosexual/straight	1343	91.73	90.32	93.15	722	94.13	92.47	95.80
gay/lesbian	25	1.71	1.04	2.37	10	1.30	0.50	2.11
bisexual	19	1.30	0.72	1.88	11	1.43	0.59	2.28
other	11	0.75	0.31	1.19	6	0.78	0.16	1.41
prefer not to say	51	3.48	2.54	4.42	N.A.	N.A.	N.A.	N.A.
don't know	7	0.48	0.12	0.83	11	1.43	0.59	2.28
refusal	8	0.55	0.17	0.92	7	0.91	0.24	1.59
total	1464				767			

Note: the category “prefer not to say” is not displayed in the IHS version as this was not one of the response option

Further research may examine whether for some specific socio-demographic groups the two protocols lead to significantly different responses on sexual identity. Also, further investigation will provide diagnostics for the ICT questions, as proposed by Glynn (2013).

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

Measurement error is a pervasive issue in social science data. It can come in different forms. For example, random error can introduce “noise” in data as people can be inconsistent when answering the same question. While this might not bias averages it can bias correlations and regression coefficients. Other types of measurement error are systematic, as such, they can bias both means and correlations. One of these is due to social desirability, the tendency of avoiding some answers in order to present oneself in a more positive light. Another example of systematic error is acquiescence, also known as “yea saying”, as people tend to agree to survey questions regardless of the content. Another example highlighted in the literature is the method effect, which indicates how the wording of question influences the answers.

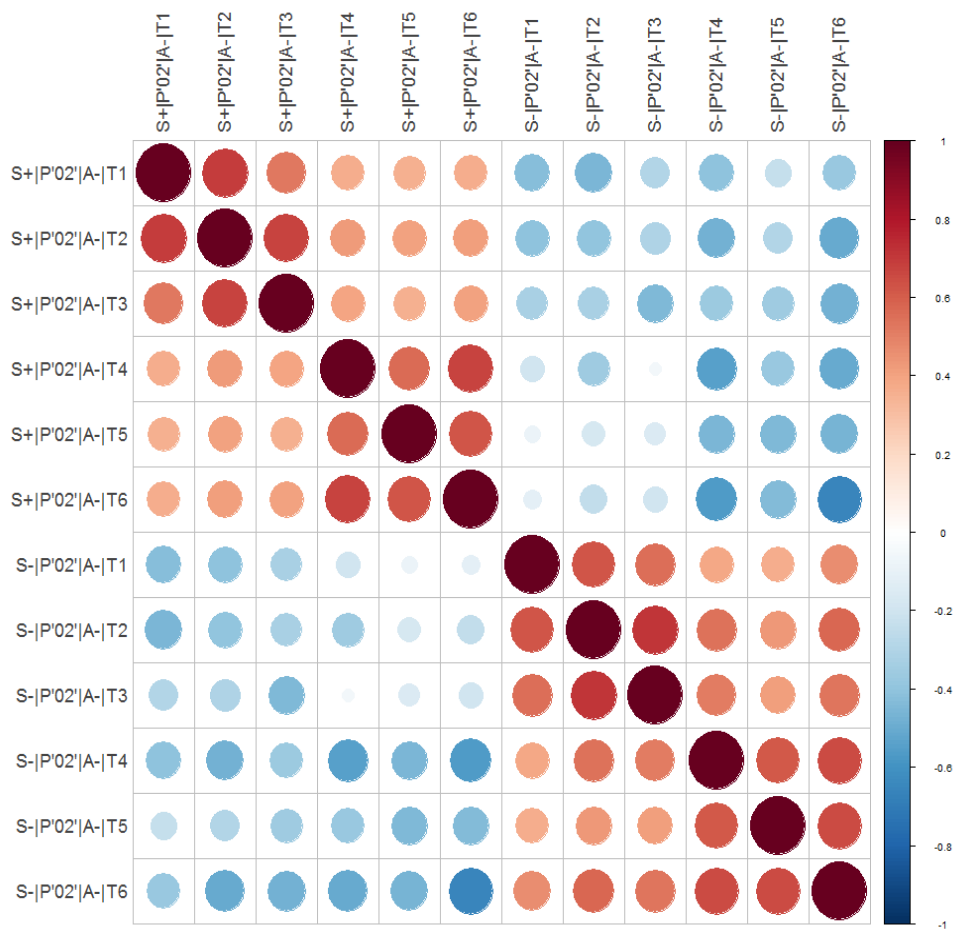
The aim of this research project is to estimate and correct for these different types of measurement error concurrently. We do this by carrying out a within person experiment where respondents receive two forms of the same questions at different points during the interview. These forms differ over 56 different randomly assigned groups in a highly fractional factorial design. In order to estimate the different types of errors we manipulate six survey questions regarding attitudes towards immigrants in three ways:

- **Number of scale points (method)**: 2 point or 11 point scale;
- **Socially desirable direction:** positively or negatively formulated items on immigration;
- **Acquiescence direction:** Agree-disagree or Disagree-agree scale.

The design of the experiment can be found in the User Guide of UKHLS-IP. Below we will present the first results from wave 8 of the Innovation Panel.

Initial results

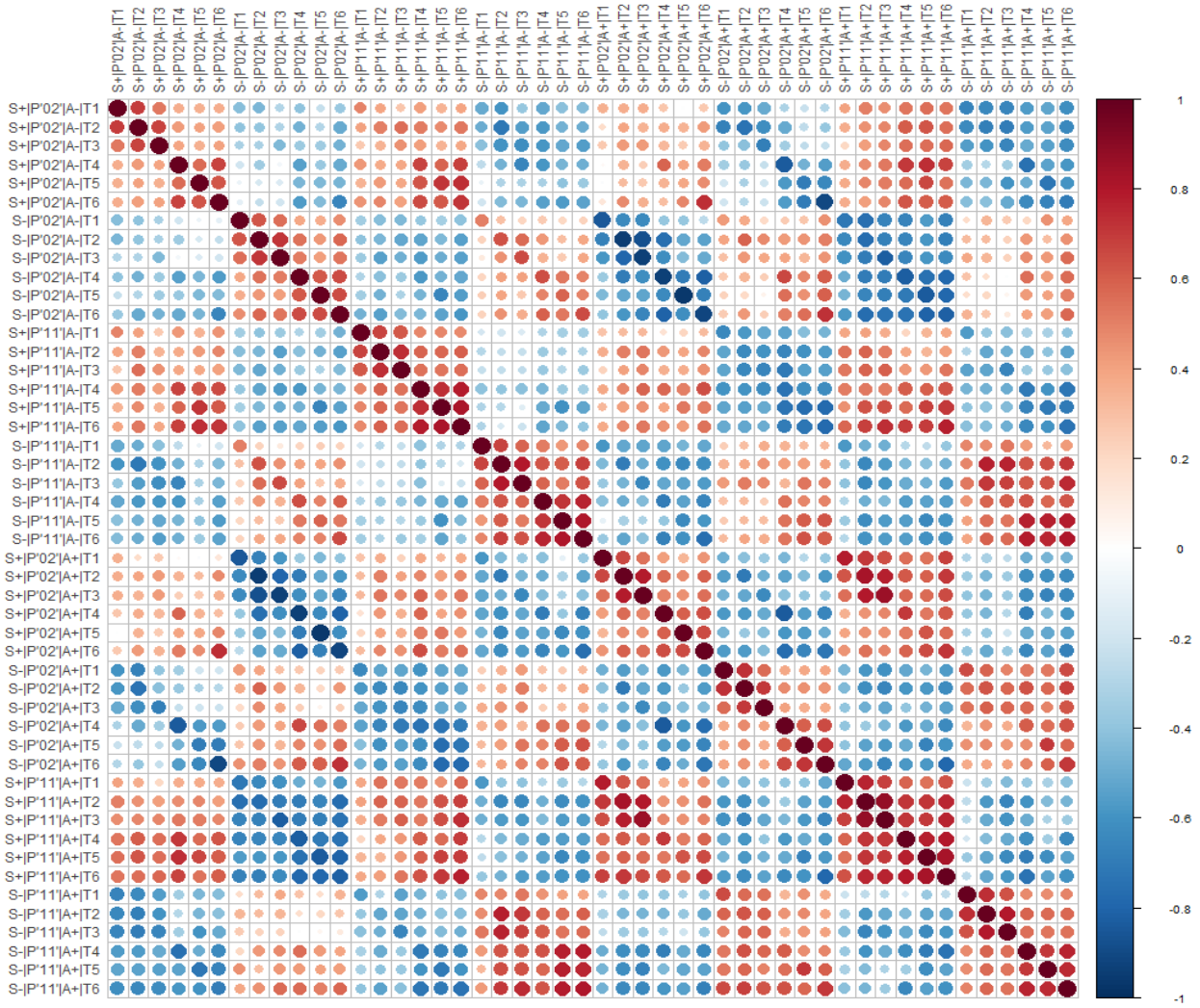
The correlation plot below represents a way to visualize the relationships between the different variables in our design. In the case below we can see the relationship between the 6 questions regarding attitudes towards immigrants (named T1-T6) when asked in two different ways.



Here we use two different wording of the survey questions in order to manipulate social desirability. As such, in the first wording we ask if respondents think there should be **fewer** people coming to the UK or if it's **bad** for the economy. In the second type of wording we reverse this, asking if there should be **more** people coming in the UK or if it's **good** for the economy. It is hoped that this will increase or decrease the direction of social desirability bias shown by the items. This manipulation is reflected in the names of the items. The first 6 items start with "S+" while the next 6 start with "S-". The other characteristics of the questions stay the same. In this case they're answers are given using a 2 point scale (P'02') using the Agree-Disagree order of the categories ("A-").

In the top left corner we can see how strong are the relationships between our 6 items. For example, we can see that the relationship between the first 3 items is stronger (larger circles and more intense red colour) than the one between the variables 4-6. This is most likely because they come originally from two different scales. If we look at the relationships between the first six rows and the last 6 columns we see that the relationship is now negative (blue colour). People are consistent with their beliefs, as such, when we reverse the wordings their answers will also change. We also see that the relationships between the six questions are different depending on the wording (compare the blue 6x6 group with the red one), indicating that the wording has an impact on the relationship between variables. We also observe that the same items have a slightly higher correlation with themselves when asked in a different way than with other variables. For example, row six with column 12 is stronger than row six with columns 7-11.

We can make this more general and look at all the 8 different ways to ask the questions for all 6 questions. This gives us a correlation plot of 48 variables. Here we see a similar pattern as before. Each new manipulation of the questions reverses the relationship with the previous one, leading to checker pattern. Within each manipulation wee that the relationships between variables change. This indicates that the wording has a strong effect on our measurement and on the correlations. Nevertheless, we can see that overall, within each 6x6 square the diagonal is stronger than the rest, meaning that each question as a strong relationship with itself even if it is asked in a different way.



This research design gives us the possibility to investigate both how systematic error impacts the means of the observed variables but also their variance. This means that we can estimate the amount of variance due to social desirability, acquiescence and method. This is important as this variance can bias analyses that use the observed survey questions. The proportion of variance can be estimated using restricted factor models, in which the loading matrices are determined by the design. This can be seen as an extension of the well-know "multitrait-multimethod" class of models (Cernat & Oberski, forthcoming; Saris & Gallhofer 2007).

In our experiment with the six questions measuring attitudes towards immigrants, the "true" or unbiased measure of attitudes towards immigrants has an unstandardized variance of 2.6 for the questions asked using the 11 point scale. In comparison to this acquiescence, or "yea saying", and the method effect, to the 11 point response scale, have a variance of approximatively 0.9. While these are a third of the true score, they nevertheless explain an important part of the observed variance. Similarly, social desirability has a significant variance (8.3) and a strong negative correlation with the true score (-0.66). This means that the people who have more negative views of immigrants are more likely to answer in a socially desirable way.

We have shown that with this design we can disentangle four different sources of errors found in the questions regarding attitudes towards immigrants in the Innovation Panel. This can be important for two main reasons. Firstly, it can inform survey designers about the biggest threats to validity and hint at possible solutions. For example, questions in the main Understanding Society survey can be modified in such a way to minimize some of these errors, such as social desirability. Secondly, our model can be used to correct for the measurement error found in the questions analysed here. So, researchers wanting to analyse attitudes towards immigrants in the Innovation Panel could do so without the biasing effects of social desirability, acquiescence, method and random error.

d. Examining the Validity of Interviewers' Ratings of Respondents' Health (Dana Garbarski, Nora Cate Schaeffer, Jennifer Dykema)

Interviewers' assessments of respondents' health (IRH) have the potential to augment significantly the power of self-rated health measures (see, e.g., Todd and Goldman 2013) in a way that is relatively inexpensive and could be incorporated in a wide variety of studies and study designs. However, a better understanding of what factors contribute to interviewers' assessments of respondents' health should be explored. This study seeks to examine the validity of interviewers' ratings of respondents' general health status with an experiment that varies when in a face-to-face interview the interviewer is asked to rate the respondent's health: at the beginning of the interview before any substantive questions are asked or at the end of interview.

We first examine whether the distribution of IRH varies depending on where in the interview the interviewer is asked to rate the respondent's health. We examine this for the 1,440 cases that were face to face survey interviews. Respondents were randomly assigned to have the interviewers rate their health at the beginning or the end of the interview. Our results show that the there is a significant difference in the distribution of interviewers' ratings of respondents' health ($p<.001$) (see Table 10). Descriptively, it appears that interviewers are more likely to select excellent health for respondents and less likely to select good health (with a bit of a difference for very good and fair health) when they assess the respondent's health later in the interview compared to early in the interview in the in-person mode.

Table 10. Interviewer Ratings of Respondent Health

	Early	Late	Total
Poor	24	27	51
	3.39	3.69	3.54
Fair	58	43	101
	8.19	5.87	7.01
Good	138	77	215
	19.49	10.52	14.93
Very good	222	208	430
	31.36	28.42	29.86
Excellent	266	377	643
	37.57	51.50	44.65
Total	708	732	1,440
	100.00	100.00	100.00

These results indicate that IRH varies depending on when interviewers are asked to assess it. However, this analysis does not indicate which version of IRH is more valid. In future analyses, the proposed experimental design will allow us to examine the validity of IRH by examining the association between IRH and other measures of respondents' health and well-being that are concurrent or precede IRH. This will allow to examine whether the predictive power of the interviewers' ratings of the respondents' health primarily stems from

interviewers' initial impressions of respondents' physical cues such as appearance and functioning, which can be gleaned in the moments leading up to the first substantive interview question, versus summarizing across the physical cues and the respondent's answers to the survey questions.

e. The impact of interesting questions on attrition (Olena Kaminska, Peter Lynn)

This experiment tests the idea that interesting questions asked towards the end of the questionnaire may improve response propensity in the following wave. The idea was tested on LISS panel before (Oudejans and Scherpenzeel, 2012) where the authors found that extra questions on politics, health or music (both tailored to participant interests and not) improved the overall feeling about the survey questions (rating them more fun and interesting) and improved long term participant retention rate.

The experiment reported here uses information from three waves of IP: wave 2, wave 7 and wave 8. At wave 2 as part of the questionnaire respondents were asked a block of specific questions on their engagement in different sport and leisure activities (see Leisure-Culture-Sport module in wave2). Our experimental treatment was implemented at wave 7, where for those people who had valid answers to the above questions we asked follow-up tailored questions. The tailored questions reflected on specific activities that respondents used to do at the time of IP2, asking whether they still engage in a particular activity and a few more details about it (see Interesting Questions module in wave 7). While the questions varied reflecting the activity we asked about, each person was only asked about one activity and at most three questions. For those who didn't have a valid response to any of the activity questions at wave 2, and those who missed wave 2 or joined the survey later we asked a more general question about their favourite TV programme (see questions tvprogreg-tvmostenjy in wave 7). Note that both refreshment samples that joined IP at waves 4 and 7 could not have been asked specific questions as they did not participate at wave 2. Thus their treatment group includes only questions on TV programs. Therefore refreshment and original samples are analysed separately. Each sample had a random assignment to two groups: control group that had no extra questions, and treatment group which received extra interesting questions towards the end of the interview (Table 11). All the analysis reported is conditional on participation at wave 7.

Table 11. Experimental group allocation within the original and refreshment samples of IP

		Original sample	Refreshment sample	Total
Control		570	577	1,147
	Specific	348	0	348
Treatment	TV	207	613	820
Total*		1,125	1190	2,315

*Total numbers reflect the number of adult full interviews in IP7 excluding those who became ineligible between IP7 and IP8 (e.g. those who died and moved out of country)

Our main interest is in attrition and whether it is affected by interesting questions asked in the previous wave. We hypothesised that respondents who are asked interesting questions may have a more positive feeling about the survey and therefore happier to respond in the future. We also expected that general TV questions would perform worse than specific questions given that the latter are tailored to specific interests of respondents.

In short we did not find any support for our main hypothesis. There were no effect on attrition of interesting questions (specific + TV) for the original sample, no effect on attrition of TV questions alone for the refreshment samples, and no effect of specific questions alone for the selected subsample of the original sample where control and treatment groups were restricted to those who would have gotten specific questions (i.e. had valid responses to wave 2 questions). See Table 12 for details.

Table 12. Comparison of response rates at wave 8 between control and treatment groups separately for the original and the refreshment samples and a subsample eligible for specific questions.

	Control	Treatment	Chi sq (1)	p-value	n
original sample (spec. + TV)	83.3	84.7	0.38	0.54	1125
refreshment sample (TV)	82.3	81.7	0.07	0.79	1190
specific q-n group (spec.)	87.2	86.5	0.08	0.77	716

While no main effect was found we further hypothesised that there may be some subgroups that may benefit from interesting questions more than others. We tested the effect of interesting questions on attrition within the subgroups of gender, age, mode of response, mode of assignment, whether respondent expects to move in the following year, weight,

whether a person is a smoker, marital status, as well as satisfaction with life, health, income and leisure.

No significant effect was observed in any of the subgroups except for the group of divorced, separated and widowed respondents for the original IP sample (Table 13). Within this group the impact of interesting questions was large: the response rate was 13.5 percentage points higher among panel participants who received interesting questions in previous wave than in control group. The effect is a result of specific questions rather than TV questions: the difference between treatment (only TV questions) and control group in refreshment samples is 1.5 percentage points (Chi-sq(1)=0.08, p=0.78); while the difference between treatment group (only specific questions) with RR=96.6 and control group with RR=82.4 among original IP subsample that was eligible for specific questions is 14.2 percentage points (Chi-sq(1)=6.39, p=0.01). Nevertheless this apparently significant finding does not withstand the multiple group test. In effect we tested the effect within 27 groups. If we are interested in p-value of 0.05, and in the context of the effect being expected in a particular direction, we could use p=0.1 as a one sided test. To correct this for multiple tests we should have a p-value of 0.0037 or lower for the result from multiple group test to be considered significant. The effect for divorced / separated / widowed category is significant at p=0.009, which is higher than 0.0037 and therefore could have been observed due to chance.

Table 13. Effect of interesting questions (specific + TV) within subgroups for the original IP sample

		Control	Treatment	Chi sq (1)	p-value	n
gender	male	84.2	84.2	<0.001	1.00	500
	female	82.7	85.1	0.67	0.41	625
age	16-30	84.3	79.0	0.86	0.35	184
	31-55	80.3	83.2	0.64	0.42	448
	56+	85.8	88.3	0.72	0.40	493
mode of response	CAPI	79.3	80.4	0.13	0.72	631
	Web	88.1	90.5	0.73	0.39	493
mode of assignment	CAPI	79.2	78.1	0.06	0.81	360
	Web+Capi	85.1	88.2	1.55	0.21	765
expecting to move	yes	69.5	79.4	1.65	0.20	127
	no	85.2	85.6	0.03	0.87	973
weight	0	77.1	80.5	0.59	0.44	407
	<=0.872	90.5	89.9	0.04	0.84	437
	>0.872	81.9	83.2	0.08	0.78	281
smoker	yes	75.0	75.0	<0.001	1.00	156
	no	84.7	86.2	0.45	0.50	969
marital status	single	81.6	82.0	0.01	0.92	291
	married / civil partner	85.7	83.8	0.42	0.52	630
	divorced / separated / widowed	77.9	91.4	6.76	0.01	197
satis with life	satisfied or neither	83.9	84.2	0.01	0.92	984
	dissatisfied	78.8	88.0	2.18	0.14	141
satis with health	satisfied or neither	82.4	84.9	0.99	0.32	881
	dissatisfied	86.7	83.9	0.38	0.54	224
satis with income	satisfied or neither	83.0	84.4	0.29	0.59	877
	dissatisfied	84.4	85.8	0.10	0.75	248
satis with leisure	satisfied or neither	82.7	87.9	1.34	0.25	876
	dissatisfied	83.5	83.8	0.01	0.90	249

Similar subgroup analysis was conducted for refreshment sample and we found no significant effect of asking TV questions in the previous wave on participation rate at the current wave for any of the subgroups.

Finally, we look at the subgroup that was or would have been assigned to TV questions in the original IP sample (i.e. those who did not have valid answers to Leisure-Culture-Sport module in wave2). Interestingly, among these we find that asking TV questions at wave 7 increases response rate at wave 8 by 14.6 percentage points (from 71.76 for control group to 86.36 for treatment group, chi-sq=4.64, p=0.03, n=151). Thus, there is some evidence that there may be a subgroup (e.g. those with no sport or leisure activities) for whom TV questions improve follow-up response rate.

Overall, with one exception, we find no effect of interesting questions asked in the previous wave on participation rate at the current wave, whether the questions are tailored to respondents' sport / culture / leisure activities or for general questions about respondents' favourite TV programmes.

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

The purpose of our experiment is to examine whether and how the direction of a response scale affects survey responses. This experiment independently varies two factors, yielding a 2*2 factorial design. The first factor manipulates the direction of response scales shown to respondents at Round 8 while holding constant other scale features. Six survey items employing two different scales are subject to this experimental manipulation, as shown in Table 14. Respondents are randomly assigned to descending scales that start with the positive/high end (e.g., completely satisfied, excellent) or ascending scales beginning with the negative/low end (such as strongly disagree, completely dissatisfied, or poor).

Table 14. 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

These six items were also included in the prior round of interview (Round 7). The second factor takes advantage of the longitudinal nature of the panel and manipulates whether or not respondents were assigned the same scale direction in the prior round as in this current round. Half of the respondents received scales in the same direction across these two waves and half received scales in different direction (for instance, some were assigned ascending scales in Round 7 and descending scales in Round 8).

We first examined the main effect of the scale direction manipulation on survey responses at Round 8, regardless of whether respondents received scales in the same direction or in the different direction in the prior round. We found that significantly more people reported from the low side when scales started from the low side than when scales started from the high side (Table 15). For instance, when the response scale starts with the negative end (“poor”), 27.4% 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 21.8%. This difference of 5.6 percentage points is statistically significant at the .05 level, confirming the presence of scale direction effect (see Table 15). Similar trends have been found in answers to the 5 satisfaction items. Significantly more reports of dissatisfaction (the low side of the scale) are found when the scale starts the low side than when the scale starts with the high side. As shown in Table 15, 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 addition, we found that the mode of data collection at Round 8 does not interact with scale direction – scale direction effect is shown in both the CAPI and the Web mode. Furthermore,

the significant main effect of scale direction still holds after controlling for mode 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).

Table 15. Scale Direction Effects Across Items

	Proportion Choosing From the Low Side				**Multivariate Model Results**	
Survey Item	Ascending Scale Starting with Low end	Descending Scale Starting with High end	Differences	p-value	Scale Direction Main Effects (Coefficients in Log Scale)	p-value
General Health	27.4%	21.8%	5.6%	0.003	-0.32	0.005
Job Satisfaction	24.0%	10.9%	13.1%	<0.0001	-0.97	<0.0001
Health Satisfaction	25.4%	19.0%	6.4%	0.0003	-0.37	0.002
Income Satisfaction	25.3%	19.3%	6.0%	0.0009	-0.35	0.004
Leisure Satisfaction	24.0%	19.1%	4.9%	0.006	-0.41	<0.0001
Overall Satisfaction	14.4%	8.9%	5.5%	<0.0001	-0.57	0.0003

We then linked Round 7 data to the Round 8 data and further investigated whether or not the observed scale direction effects are affected by respondent experience with scales. Specifically, we wanted to examine whether or not scale direction effects are stronger for those who received scales in the same direction across the two interview administrations. We fit logistic regression models predicting the likelihood for respondents to respond in the low side of the scale at Round 8 by the scale direction manipulation at Round 8, across-wave scale direction manipulation, an indicator that respondent chose from the low side in Round 7, interaction between Round 8 scale direction manipulation and across-wave scale direction manipulation, interaction between Round 8 scale direction manipulation and Round 7 responses, mode of data collection in Round 8, and demographic characteristics such as age, gender, marital status, and employment status.

Table 15. Partial Logistic Regression Results

	Partial Logistic Regression Results					
	General Health	Job Satisfaction	Health Satisfaction	Income Satisfaction	Leisure Satisfaction	Overall Satisfaction
Scale Starting with High End	**-0.92**	**-1.42**	**-0.93**	**-0.59**	**-0.79**	**-0.92**
Same Scale Direction In both Rounds	-0.23	0.31	-0.11	0.14	-0.28	0.01
Responded from Low Side in Round 7	**2.53**	**0.90**	**1.82**	**1.65**	**2.00**	**1.85**
Scale Direction * Same Scale	**0.69**	-0.04	**0.58**	0.40	*0.50*	0.37
Scale Direction * Round 7 Responding	*0.51*	**0.98**	*0.52*	0.05	0.37	0.41
CAPI in Round 8	-0.08	-0.10	-0.03	0.06	0.07	-0.06

Note: Coefficients for demographic characteristics are not shown. Coefficient estimates are in log scale. Bold estimates are statistically significant at p=0.05 level and italicized estimates are statistically significant at p=0.10 level.

Shown in Table 15 are partial logistics regression results. After controlling for demographic differences, mode of data collection at Round 8, and other experimental manipulations, scale direction has a significant effect on respondents' likelihood to choose from the low side; scales starting with the high end reduce the likelihood to choose from the low side than scales starting with the low end. Whether or not respondents received the same scale at both rounds doesn't seem to matter. However, respondents who chose from the low side of a scale in Round 7 are more likely to choose from that side again. The interaction between Round 8 scale direction manipulation and the across-wave scale manipulation is statistically significant for two of the models and marginally significant for one model. The positive interaction indicates that scale direction effect is observed regardless of scale direction manipulation at Round 7 and that the group who were assigned scales beginning with the low end at both rounds produced most reports from the low side of the scale at Round 8.

The interaction between Round 8 scale direction manipulation and respondents' answers at Round 7 is statistically significant for one model and marginally significant for two models. The positive interaction suggests that scale direction effect is observed for both those who chose from the low side at Round 7 and also those who chose from the high side at Round 7 and that the highest proportion reporting from the low side of the scale at Round 8 comes

from respondents who were given a scale starting with the low side at Round 8 and who also chose from the low side at Round 7 and the lowest proportion comes from the group that were assigned a scale starting with the high side at Round 8 and who chose from the high side at Round 7.

Our preliminary results demonstrate that scale direction affects survey answers by pushing answers to the start of the scale regardless of mode. Respondents who reported from the low side of a scale at an earlier round are especially more likely to choose from that side again at Round 8, especially when the same ascending scale was also assigned at Round 7. These results have a great implication for survey researchers and the survey community.

g. Respondent Incentives (Peter Lynn)

Most sample members received the same incentive at IP8 as they had done at IP7. In consequence, there were again three experimental groups amongst the continuing mixed mode sample (£10 unconditional, with or without an additional £20 conditional on participation online, £30 unconditional), three experimental groups amongst the IP7 refreshment sample (£10, £20 or £30, unconditional), and no experimentation amongst the continuing CAPI-only sample (£10 unconditional).

A small subset of the IP7 mixed mode sample had been identified, after three mixed mode waves, as having a very low propensity to respond online. These sample members were switched to CAPI-only at IP8 and all were administered a £10 unconditional incentive. For some, this represented a reduction in their incentive from £30 at the previous waves. These were the only sample members whose incentive changed at IP8. Amongst the low online propensity sample members switched to CAPI-only, response rate at IP8 was 85% for those who received a £10 unconditional incentive at IP7 (and for whom the IP8 incentive therefore represented no change) and 75% for those who received a £30 unconditional incentive at IP7 (and for whom the IP8 incentive therefore represented a substantial reduction). Though this may appear to suggest that reducing the level of incentive harms response rate, the difference is not statistically significant due to the small sample sizes involved (fewer than 50 sample members in each of the two groups compared, P=0.32).

As the incentive treatments were kept unchanged for three waves in the continuing mixed mode sample, it is possible to test the effect of different incentive levels on sample retention over three waves. We find (table 16) that sample retention is significantly, and substantially, higher with a £30 incentive than with a £10 incentive. Of all adult mixed mode sample members issued to the field at IP6 and not known to have died or moved abroad by the time of IP8, the proportion still responding at IP8 was 71.1% amongst those sent £30 at each wave, compared to 55.6% of those sent £10 (and 61.5% of those sent £10 and offered a further £20 conditional on all household members responding online): $P = 0.0003$. The effect size was even larger amongst the original sample (71.2% retention with £30 incentive vs. 51.9% with £10; 63.2% with £10 plus £20 conditional; $P = 0.0006$), but was not significant amongst the IP4 refreshment sample (70.8% with £30 vs. 63.4% with £10; P = 0.28).

Table 16: Sample retention rates in mixed mode sample, by incentive

	Incentive		
Most recent wave responded	£10U %	£10U+£20C %	£30U %
IP8	55.6	61.5	71.1
IP7	7.7	9.3	5.8
IP6	12.5	11.3	9.8
IP5 or earlier	24.3	17.9	13.3
n	585	603	602

Design-based $F(5.20, 312.05) = 4.635$, $P = 0.0003$; *n* is number of persons aged 16 or over issued to the field at IP6 and believed to still be eligible at the time of IP8 (i.e. not known to have become ineligible)

References

Cernat, A., & Oberski, D. L.. (submitted). Extending the within-persons experimental design: the multitrait-multierror (MTME) approach. In Lavrakas, P. J. (Ed.), In Experimental methods in survey research New York: John Wiley & Sons.

Droitcour, Judith, Rachel A. Caspar, Michael L. Hubbard, Teresa L. Parsely, Wendy Visscher, and Trena M. Ezzati. (1991). "The Item Count Technique as a Method of Indirect Questioning: A Review of its Development and a Case Study Application". In Paul P. Biemer, Robert Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman (Eds.), *Measurement Errors in Surveys* (pp. 185–210). New York: Wiley

Glynn, Adam M. (2013) What Can We Learn with Statistical Truth Serum? Design and Analysis of the List Experiment. *Public Opinion Quarterly* 77:159-172

Saris, W., & Gallhofer, I. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research* (1st ed.). Wiley-Interscience.

Tourangeau, Roger and Ting Yan. (2007). Sensitive Questions in Surveys. *Psychological Bulletin* 133 (5): 859–883

Tourangeau Roger, Kenneth Rasinski, and Lance J. Rips. (2000). *The Psychology of Survey Response*. Cambridge University Press

Oudejans, Marije and Annette Scherpenzeel (2012) "Especially For You: Motivating Respondents in an Internet Panel by Offering Tailored Questions." Paper Presented at the 8th International Conference on Social Science Methodology (RC33), Sydney.

Todd, Megan A., and Noreen Goldman. 2013. "Do Interviewer and Physician Health Ratings Predict Mortality?: A Comparison with Self-Rated Health." *Epidemiology* 24(6):913–20.