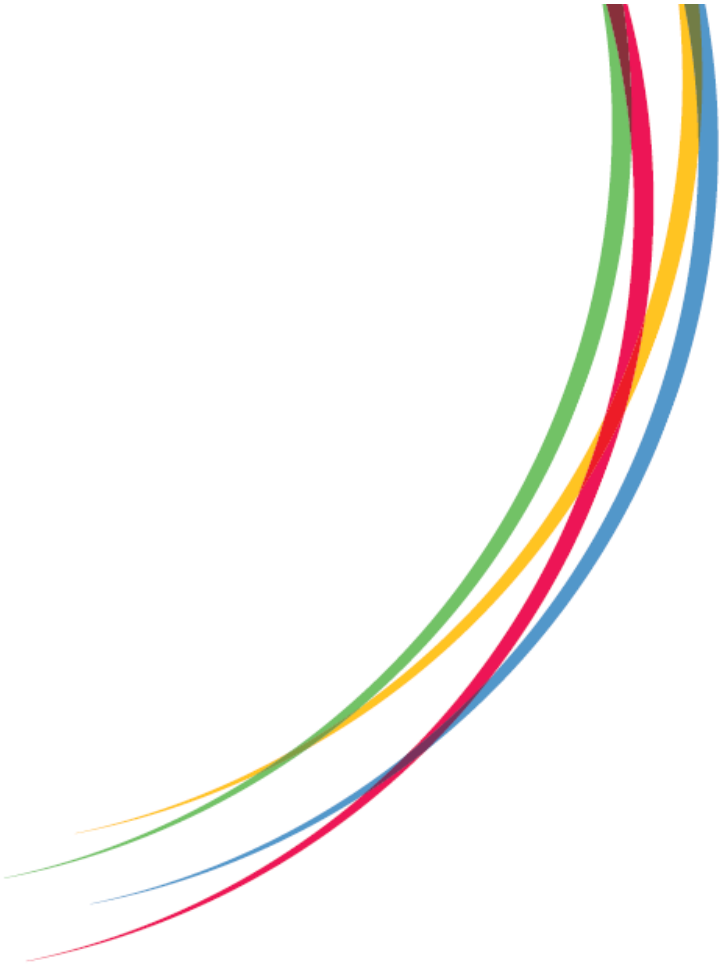


Understanding Society

Working Paper Series

No. 2017-07

July 2017

Understanding Society Innovation Panel Wave 9: Results from Methodological Experiments

Alessandra Gaia[1] (ed.)

Contributors: Michaela Benzeval[1], Annamaria Bianchi[2], Mike Brewer[3], Jonathan Burton[1], Alexandru Cernat[4], Mathew Creighton[5], Thomas Crossley[6], Adeline Delavande[1], Paul Fisher[1], Alessandra Gaia[1], Annette Jäckle[1], Amaney Jamal[7], Daniel Oberski[8], Frank Popham[9], Elise Whitley[9], Joachim Winter[10], Basit Zafar[11]

[1] Institute of Social and Economic Research (University of Essex), [2] University of Bergamo, [3] Institute of Social and Economic Research (University of Essex) and Institute for Fiscal Studies, [4] University of Manchester, [5] University College Dublin, [6] University of Essex and Institute for Fiscal Studies, [7] Princeton University, [8] Utrecht University, [9] University of Glasgow, [10] University of Munich, [11] Federal Reserve Bank of New York

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 9 of the Innovation Panel in the spring and summer of 2016.

In IP9, as with prior waves, there were methodological experiment involving the value of respondent incentives and mixed-mode data collection. Additionally, IP9 included a series of experiments and innovations to improve the measurement of household finances. Further experiments analyse how targeted survey email invitations influence response rates; examine the measurement of attitudes on sensitive issues with a technique using item counts; analyse subjective expectations about the returns to schooling and the decision to go to university; experiment on different methods to measure what the general population regard as “successful ageing”; test how the presentation of response options impact estimates of satisfaction; and take multiple measurements to better understand attitudes and the impact of how scales are presented.

Understanding Society Innovation Panel Wave 9:

Results from Methodological Experiments

Alessandra Gaia[1] (ed.)

Contributors: Michaela Benzeval[1], Annamaria Bianchi[2], Mike Brewer[3], Jonathan Burton[1], Alexandru Cernat[4], Mathew Creighton[5], Thomas Crossley[6], Adeline Delavande[1], Paul Fisher[1], Alessandra Gaia[1], Annette Jäckle[1], Amaney Jamal[7], Daniel Oberski[8], Frank Popham[9], Elise Whitley[9], Joachim Winter[10], Basit Zafar[11]

[1] Institute of Social and Economic Research (University of Essex), [2] University of Bergamo, [3] Institute of Social and Economic Research (University of Essex) and Institute for Fiscal Studies, [4] University of Manchester, [5] University College Dublin, [6] University of Essex and Institute for Fiscal Studies, [7] Princeton University, [8] Utrecht University, [9] University of Glasgow, [10] University of Munich, [11] Federal Reserve Bank of New York

Abstract

This paper presents some preliminary findings from Wave 9 of the Innovation Panel (IP9) of *Understanding Society*: The UK Household Longitudinal Study. *Understanding Society* is a major panel survey in the UK. IP9 included experiments on the use mixed mode data collection, the value of respondent incentives, targeted timing of email invitations, measurement of household finances, subjective expectations about returns to schooling, people's assessment of what constitutes "successful ageing", format of response options, use of multiple measurements to improve measurement of attitudes, and measurement of sensitive topics. This paper describes the design of IP9, 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: Alessandra Gaia (agaia@essex.ac.uk) Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK.

Acknowledgements: The *Understanding Society* Innovation Panel is funded by the UK Economic and Social Research Council (ES/N00812X/1). The work on "Reconciling household income and spending" and on "Improving household survey measures of income" was funded by the ESRC Transformative Research scheme and the National Centre for Research Methods (ES/N006534/1). Elise Whitley and Frank Popham received funding from the UK Medical Research Council (MC_UU_12017/13) and the Scottish Government Chief Scientist Office (SPHSU13).

Table of Contents

1. Introduction

This paper presents early findings from the ninth wave of the Innovation Panel (IP9) of *Understanding Society*: The UK Household Longitudinal Study (UKHLS). *Understanding Society* is a major panel survey for the UK. The first seven waves of data collection on the main sample have been completed, and eight and nine waves are currently in the field. The data from the first six waves of the main samples are available from the UK Data Service, and the seventh will be available towards the end of 2017. 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 nine 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 about 1,500 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. Fieldwork for IP5 began in March 2012 and for IP6 in March 2013. For IP7 fieldwork began in June 2014, adding 488 responding households as a refreshment sample. IP8 fieldwork started in summer 2015, and IP9 in May 2016. Working Papers which cover the experimentation carried out in all nine innovation panels are available from the *Understanding Society* website.[2] This paper describes the design of IP9, the experiments carried and some preliminary findings from early analysis of the data. Section 2 outlines the main design

[1] http://discover.ukdataservice.ac.uk/series/?sn=2000053
[2] https://www.understandingsociety.ac.uk/research/publications/working-papers

features of *Understanding Society*. Section 3 describes the design and implementation of IP9. Section 4 then reports on the experiments carried at IP9.

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 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 Kantar Public (formerly known as TNS BMRB), and wave 9 is being carried out by Kantar Public in collaboration with NatCen[3]. *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[4]. *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 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-

[3] Waves 1 to 6 of the Innovation Panel were implemented by NatCen; waves 7 to 9 by Kantar Public; wave 10 is being implemented in collaboration by NatCen and Kantar Public.

[4] The Innovation Panel sample includes Great Britain only.

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 six waves, surveys of continuing sample members were interviewer-administered.
At wave 7, adults in households that had not participated at wave 6 were first invited to take part online; then, non-respondents were issued to face-to-face interviewers.

At Wave 8 a random 20% of households were assigned to the CAPI-only design. Of the remaining 80%, 20% of households, which were identified as having the lowest likelihood of responding in the mixed-mode design, were assigned to face-to-face interviews. The remaining households were randomly allocated to either a web-first protocol or a face-to-face first protocol. For more information see the User Guide (https://www.understandingsociety.ac.uk/documentation/mainstage).

In between each wave of data collection, adult 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. The contact procedures, incentives, and tracing strategies vary by mode allocation; more details are available in the Understanding Society User Guide. 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). From Wave 2, the BHPS sample was incorporated into the *Understanding Society* sample. The BHPS sample is interviewed in the first half of the fieldwork period for each wave.

3. Innovation Panel Wave 9: Design

IP9 employed a mixed-mode design, which started in IP5. In IP5, a random selection of two-thirds of households was allocated to a mixed-mode design with web followed by face-to-face interviews (“web-first” group); the remaining households were allocated to a single mode face-to-face design (“F2F-first” group). In IP6 to IP9, towards the end of fieldwork, a “mop-up” phase was added: all non-responding adults in both experimental groups were invited to complete the interview online or by telephone.

The sample allocation implemented at IP5 has been maintained at each subsequent wave, with only one exception: from IP8 onwards a small subgroup of households with a very low propensity to respond via the web were switched from the "web first" group to the "face-to-face first" group. The IP7 refreshment sample units were initially all allocated to the F2F interviewing only; at IP8, they were allocated to the "face-to-face first" group; at IP9 a random 2/3 were allocated to the "web first group" and the remainder to the "face-to-face first group".

For more details on the mixed-mode design and fieldwork see the IP User Guide[5] and the technical report[6].

Call for experiments

IP9 was the seventh 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 16th February 2015 with a deadline of 13th April 2015. Nine proposals were received with two being accepted. In addition to the two new experiments, four were carried over from previous waves. The nine 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 of *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 proposed by the leadership team. These core experiments included the mixed-mode design, the main incentives experiment, the experiments on household finances, and the experiment on the visual representation of satisfaction scales.

Sample

There were three samples issued at IP9: 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 IP8, plus some households which had not responded at IP8. 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 751 original

[5] IP User Guide: https://www.understandingsociety.ac.uk/documentation/innovation-panel

[6] IP technical report: https://www.understandingsociety.ac.uk/documentation/innovation-panel/technical-reports

sample households, 347 IP4 refreshment sample households and 453 IP7 refreshment sample households issued at IP9. There were 1,551 total sample households issued at IP9. All of the households were originally selected from the Postcode Address File (PAF) using the same methods.[7]

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 IP9, with only some exceptions: as noted, some households that were deemed to have a very low propensity to respond via were assigned directly to CAPI to begin fieldwork. Sample members would be approached by letter and email (where possible) to complete their interview on-line. At wave 9, approximately two thirds of the IP7 refreshment sample were all allocated a "web first" design, and the remaining households were allocated to a "face-to-face first" design. The table below shows the allocation to mode design by sample type for those included in the issued original sample and IP4 and IP7 refreshment samples in IP9.

Table 1: Allocation to mode design by sample type

	Original Sample	IP4 Refreshment Sample	IP7 Refreshment Sample	Total
F2F first	273 36.4%	136 39.2%	152 33.6%	561 36.2%
Web first	478 63.7%	211 60.8%	301 66.5%	990 63.8%
Total	751	347	453	1,551

Questionnaire design

The questionnaire at IP9 followed the standard format used in the previous Innovation Panels as well as the main-stage of *Understanding Society*. The questionnaires used at IP9 are available from the *Understanding Society* website.[8] The interview included the following sections with the corresponding target times for each:

- Household roster and household questionnaire: average 15

[7] 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

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

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.[9] 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

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

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.

Response rates

This section sets out the response rates for IP9 as a whole. The issued sample at the ninth wave consisted of 1,399 households that had responded to IP8 and 152 households that had not responded at IP8. Table 2 displays the household-level response at IP9 for the original and IP4 refreshment samples by “face-to-face first” and “web first” design, 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, IP4 and IP7 Refreshment samples, IP9 (adults only)

	Original Sample		IP4 Refreshment Sample		IP7 Refreshment Sample		Combined		Total
Household RR	F2F first	Web first	F2F first	Web first	F2F first	Web first	F2F first	Web first	
Complete HH	61.7%	67.5%	65.4%	72.6%	56.4%	63.5%	61.2%	67.4%	65.1%
	163	*314*	*87*	*148*	*84*	*186*	*334*	*648*	982
Partial HH	22.0%	18.1%	18.8%	18.1%	21.5%	20.1%	21.1%	18.7%	19.6%
	58	*84*	*25*	*37*	*32*	*59*	*115*	*180*	295
Total Responding HH	83.7%	85.6%	84.2%	90.7%	77.9%	83.6%	82.2%	86.1%	84.7%
	221	*398*	*112*	*185*	*116*	*245*	*449*	*828*	*1,277*
Nonresponding HH	16.3%	14.4%	15.8%	9.3%	22.2%	16.4%	17.8%	13.9%	15.3%
	43	*67*	*21*	*19*	*33*	*48*	*97*	*134*	*231*
Total HH	**264**	**465**	**133**	**204**	**149**	**293**	**546**	**962**	**1,508**
Conditional Individual RR	F2F first	Web first	F2F first	Web first	F2F first	Web first	F2F first	Web first	
Responding individuals	82.8%	87.0%	82.1%	87.8%	82.7%	85.9%	82.6%	86.9%	14.6%
	346	*724*	*179*	*325*	*181*	*419*	*706*	*1,468*	*371*
Nonresponding individuals	17.2%	13.0%	17.9%	12.2%	17.4%	14.1%	17.4%	13.1%	85.4%
	72	*108*	*39*	*45*	*38*	*69*	*149*	*222*	*2,174*
Total Ind.	**418**	**832**	**218**	**370**	**219**	**488**	**855**	**1,690**	**2,545**

There were 1,277 interviewed households from the continuing samples, for an 84.7% overall household response rate. Within these households, 2,174 people were interviewed, for a conditional individual response rate of 76.8%.

Given the mixed-mode design used at IP9, not all individuals responded in the same mode. Further, at IP9 the mop-up period was again used, where non-responding units in all 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 and total overall at IP9 including the mop-up phase.

Table 3. Mode of Response, IP9

	Original Sample		IP4 Refreshment Sample		IP7 Refreshment Sample		Combined		Total
Responding Mode	F2F first	web first	F2F first	web first	F2F first	web first	F2F first	web first	
Face-to-Face	90.5%	22.4%	93.9%	19.1%	93.4%	34.8%	92.1%	25.2%	46.9%
	313	*162*	*168*	*62*	*169*	*146*	*650*	*370*	*1,020*
Web	8.4%	75.8%	3.9%	80.0%	5.0%	64.2%	6.4%	73.4%	51.7%
	29	*549*	*7*	*260*	*9*	*269*	*45*	*1,078*	*1,123*
Telephone	1.2%	1.8%	2.2%	0.9%	1.7%	1.0%	1.6%	1.4%	1.4%
	4	*13*	*4*	*3*	*3*	*4*	*11*	*20*	*31*
Total Ind.	**346**	**724**	**179**	**325**	**181**	**419**	**706**	**1,468**	**2,174**

Table 4. Device Used, Web Respondents, IP9

	IP9 Web Respondents
PC/Laptop	65.5% *735*
Large Tablet	20.9% *235*
Small/Medium Tablet	6.0% *67*
Smartphone	7.4% *83*
Other	0.3% *3*
Total Web Respondents	**1123**

As at IP8, also at IP9 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 – IP9 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 IP9 is presented in Table 4.

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 is the first for IP7.

Table 6 presents the longitudinal individual re-interview rates for the original sample (for IP2-IP9), the IP4 refreshment sample (for IP5-IP9), and IP7 refreshment sample (for IP9), conditional on being eligible in the given wave. For each cell, the percent is reported above the number of individuals the percent represents, in italics.

Table 5. Longitudinal re-interview rates

	IP2	IP3	IP4	IP5	IP6	IP7	IP8	IP9
Original Sample	69.3% *1654*	60.6% *1442*	54.7% *1270*	45.9% *1095*	45.9% *1100*	38.4% *917*	36.2% *867*	35.8% *814*
IP4 Refreshment Sample	-	-	-	82.0% *586*	76.8% *554*	62.1% *447*	58.8% *423*	58.7% *396*
IP7 Refreshment Sample	-	-	-	-	-	-	79.2% *520*	82.7% *487*

As with any longitudinal study, there has been attrition at each wave, decreasing the overall numbers for each sample. At IP9, 814 individuals from the original sample who responded at IP1 were successfully interviewed, representing a 35.8% re-interview rate. For the IP4 refreshment sample, the IP7 was their fifth wave and 396 responded, for a 58.7% re-interview rate. IP8 was the second wave for the IP7 refreshment sample, with 487 responses for a 82.7% re-interview rate.

4. Experimentation in IP9

There were a number of experiments carried on IP9 covering both fieldwork procedures and measurement in the questionnaire. There were some new experiments and some which were the continuations of experiments carried at previous waves of the IP. This section outlines the experiments carried at IP9, 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 of each sub-section below are given in the heading.

(1) Reconciling Household Income and Spending (Mike Brewer, Jon Burton, Thomas F. Crossley, Paul Fisher, Alessandra Gaia, Annette Jäckle, and Joachim Winter)

Objectives

As an accounting identity, income minus expenditure must equal active saving (that is, additions to or withdrawals from net wealth). This accounting identity is implemented in the household sectors accounts within the national accounts – in the "Use of Disposable Income" account, resources (disposable income) are balanced by uses (household expenditure and saving, where the latter can be negative) – but it should hold for individual households too. The objective of this experiment was to explore how to implement this identity in a multi-purpose, mixed-mode household survey.

Capturing all of household or benefit unit income, consumption and active saving (or additions and subtractions from wealth) in a single survey would allow for the calculation of covariances between these components, and these covariances can be critical to distinguishing between alternative explanations for household behaviour. In addition, evidence on the use of a "balance edit" in a Canadian survey of household spending (Brzozowski & Crossley 2011) and in experiments conducted by the Bureau of Labour Statistics (BLS) (Fricker *et al.*, 2015) suggests that it could improve data quality. The Townsend Thai Monthly Survey (Samphantharak & Townsend 2010) has pioneered the collection of information about all components of the household accounting identify in a developing country, and demonstrated its advantages, both in terms of data quality and substantive research possibilities.

Description of the Experiment

Implementation Details

In Understanding Society, income and wealth are collected as part of the individual interviews. One difficulty is the possibility of double-counting, that is, that the same income source may be collected from more than one individual in the household and so counted twice when deriving the income of the whole household. This is especially problematic where respondents are living as a couple and may have highly inter-related finances. Spending is collected at the household level in Understanding Society. However, it can be very difficult for a single respondent to report the spending of a complete household, particularly a complex household with a number of autonomous economic agents (such as adult children living with parents, or "sharers"). As a consequence, it was decided to collect and reconcile information on income, spending and change in net financial assets at the level of the benefit unit. A Benefit Unit (BU) is defined as a single adult without a co-resident partner or spouse plus any dependent children, or a co-resident couple (whether married or cohabiting) plus any dependent children. This concept is used in other UK household surveys – notably the Family Resources Survey – and it corresponds to the unit of assessment for most means-tested benefits and tax credits in the UK. Additional adults within a household (such as adult siblings, non-dependent children, or unrelated adults living together but not in a relationship) are a separate benefit unit. We therefore refer this set of questions as the "Benefit Unit Module".

Where respondents were living with a spouse or partner, the Benefit Unit Module was designed to be asked of both members of the couple together. This could be done only if both partners agreed for the financial information they had given in their individual interview to be shared with their partner. If both partners agreed, then the questions in the Benefit Unit Module were asked at the end of the second person's interview. If either adult in a couple did not give permission, then these questions were not asked. For face-to-face interviews, when both partners agreed, the interviewer coded which respondents answered these questions (either one of the partners, or both of them together). For online interviews, respondents living with a spouse or partner were asked to complete these questions together if possible. They were also asked to confirm if both of them or only one of them had answered the questions. Single adult Benefit Units were asked the question at the end of their individual interview.

The Benefit Unit was not asked to: (a) proxy respondents, (b) sample members completing the survey by telephone, (c) sample members which form part of a Benefit Unit where at least one of the partners hasn't agreed to take part (d) respondents aged 16-19 and in full time education and living with their parents (e) sample members which form part of a Benefit Unit were the first partner was interviewed face-to-face and the second was interviewed by web.

The intent of the Benefit Unit Module was to collect data on the Benefit Unit's net income, spending and the change in net financial assets, all over the same period of time. For net income, the script calculated a Benefit Unit total of monthly net income from the amounts already reported in the individual interviews. Benefit Units were then asked to confirm or correct that amount. Monthly total spending was elicited using a single question, based in part on an experiment in the sixth wave of the Innovation Panel (see Al Baghal *et al*., 2014). The method of eliciting changes in net financial assets was varied experimentally, as described in the next subsection.

With Benefit Unit-level data on total monthly net income, total spending, and additions and subtractions from wealth, the script calculated the difference between 'Incoming Money' (income plus any increase in borrowing or drawdown of savings) and 'Outgoing Money' (spending plus any reductions in debts or increases in savings). We refer to this difference as the "balance", although according to the accounting identity discussed above it should be identically zero. Where the amounts did not balance (i.e. where 'Incoming Money' was not equal to 'Outgoing Money'), respondents were asked to check the amounts and reconcile any differences. Respondents were not forced to make the amounts balance to zero.

Experimental variation

Households were randomly allocated to one of two ways of eliciting changes in net financial assets.[10] Half were asked the 'gross flows' model, in which outgoing money (spending, new savings and debts that have been paid off) was subtracted from incoming money (income, increases in debts and withdrawals from savings). Benefit units in the other

[10] The randomization is at the household level, so that where multiple responding Benefit Units in the same household receive the same treatment. This should be accounted for in inference.

half of households were asked the 'net flows' model. They were asked about net changes in a range of financial assets (including debts), which were then aggregated to give a change in net financial assets. The difference between income and spending was then compared to the aggregate net change in the benefit unit's financial assets.

Interaction with other experiments

This experimental variation has important interactions with at least two others. First, the degree of imbalance and the revision behaviour of responding benefit units may depend on whether individuals were exposed to the income reconciliation screen in experiment "b. Improving Household Survey Measures of Income". Second, the Benefit Unit Module may work differently in web and face-to-face modes. Randomisation in this experiment was stratified with respect to allocated mode.

Key research questions

This experiment seeks to address a number of key research questions. Here we present preliminary results on two:

i) *Participation:* Will individual members of couples consent to a benefit-unit level household finance module (and hence share their income responses with each other)?

ii) *Balance and reconciliation:* Do benefit unit level reports of income, spending and additions and changes in net financial assets balance? If confronted with imbalance, do responding benefit units revise their responses? What kinds of changes do they make? How many more benefit units are brought into balance through the revision process?

Additional questions for future research include: Do responding benefit units revise their benefit unit level net income from the total derived from individual data? Does the balancing process lead to an improvement in data quality? Which method of collecting additions and subtractions from wealth gives better data? Does the effect of balancing vary by survey mode, and/or by whether the individual interview included the income summary screen?

Results

Data availability and consent

Overall, there is usable data from the Benefit Unit Module for 1,044 BUs, or 63 percent of all benefit units in the 1,277 responding households. Table 6 shows the reasons why not all benefit units in respondent households were asked the Benefit Unit Module. The requirement

for both adults in a couple to consent is the biggest reason why the sample that is asked the Benefit Unit Module questions is slanted towards single adults. Of the 1,174 adults in couples asked for consent, 77% gave consent, but only in 66% of couples did both adults give consent. There was little difference in completion or consent rates by survey mode (other than those that used the phone were not eligible for the module at all).[11]

Table 6. Derivation of final analysis sample for Benefit Unit Module

Sample	N adults aged 16+ and not in F.T. education	N Benefit Units	*Of which, share single adults*	N households	*Of which, share on web*
Respondent households	2,421	1,646	0.516	1,277	0.649
…where all adults provided individual interviews	2,014	1,373	0.533	1,146	0.646
…and no interviews in the BU were by telephone, or proxies	1,886	1,290	0.538	1,096	0.649
…and, if adults in couples used different modes, the second one was not web	1,868	1,281	0.542	1,090	0.651
…and both adults in a couple consented to share information with each other	1,471	1,081	0.639	930	0.628
…and who appear in i_bufinance.dta	1,399	1,044	0.660	895	0.636

Note: "F.T." indicates full time education.

Benefit Unit budget reconciliation

As described earlier, the aim of the BU Module was to collect information on BU level net income, spending and the change in net financial assets, and then to prompt Benefit Units if the combined responses to these were inconsistent. In the analysis that follows, we limit attention to the 838 BUs (out of 1,044) that reported non-zero values of both income

[11] There are 36 BUs that gave consent but do not appear in i_bufinance.dta, 35 of which are couples, and there is also 1 BU which gave consent and appears in i_bufinance.dta but where the type of benefit unit is not coded consistently across the survey.

and spending. We do this because it is not always clear to what extent zero values are genuine, or a default value for BUs that could not report a value.[12]

We focus here on whether BUs were able to report consistent information on their finances, that is, whether the BU's finances were "in balance", which is defined as having an absolute value balance level that is less than 10% of the average of income and spending, i.e. if $|(income - spending - net\ increase\ in\ assets)| < 0.1\left(\frac{income+spending}{2}\right)$. Table 7 reports this, splitting BUs by their experimental allocation.

Table 7. Number of Benefit Units who reported non-zero values of income and spending who report to be in balance before and after household budget reconciliation, by Treatment

	Gross flows	Net flows
In balance before	90	106
In balance after	155	162
Total	402	436
Of those initially out of balance:		
balance changed	*140*	*119*
where income changed	*47*	*52*
where spending changed	*75*	*59*
where "change in assets" changed	*76*	*72*
abs(balance) fell	*133*	*107*
abs(balance) rose	*7*	*12*
Total	*312*	*330*

Note: in balance defined as: $|(income - spending - net\ increase\ in\ assets)| < 0.1\left(\frac{income+spending}{2}\right)$

[12] In particular, we can observe that 15% of BUs either did not give an answer to the question on monthly spending, or gave an answer of £0, which we view as implausible. 11% of BUs in the [gross flows] allocation, and 5% of BUs in the [net flows] allocation, refused to answer, or did not know the answers to, questions on how their debts and savings balances had changed. Before calculating the final balance, the CAPI and CAWI routines set any incomplete or missing values for spending or for changes in assets to zero. There are no un-usable values of income in the data-set at this stage. Net income is £0 in 9% of BUs, which is a far greater fraction than suggested by other households surveys, suggesting that some might reflect item non-response; on the other hand, all respondents had already had the opportunity to review and amend their income.

Overall, 23% of BUs were in balance before the reconciliation, and 38% after. Table 7 shows that the "gross flows" method had slightly fewer BUs initially in balance, but then slightly more in balance after the reconciliation. The "change in net financial assets" is the most likely concept to be adjusted, followed by spending, and then by income; this is unsurprising given that participants have already had at least 1 chance to review their income, and that the question about spending was a one-shot. In general, these findings – that reconciliation reduces, but far from eliminates, budget imbalance is consistent with a small-scale lab study reported in Fricker *et al* (2015).

Conclusion

Preliminary analysis of this experiment shows, first, that it is feasible to collect, at the Benefit Unit levels, all of the elements of the accounting identity: net income minus spending equals change in net financial assets. Most adults in couples were willing to participate in the benefit module, although the requirement that both adults in a two-adult Benefit Unit consent meant that the participation rate was higher for single-adult Benefit Units.

Second, our analysis shows that while the reconciliation process increased the number of Benefit Units in balance by 65% (121/196), the low initial rate of balance means that more than half of household remain out of balance, post reconciliation.

(2) Improving Household Survey Measures of Income (Paul Fisher, Alessandra Gaia, Mike Brewer, Jon Burton, Thomas F. Crossley, Annette Jäckle, and Joachim Winter)

Objectives

Income data collected as part of household surveys is critical for the study of material living standards; however, it is known to suffer from misreporting. We implemented a series of experiments designed to reduce misreporting; specifically, we tested the implementation of an end-of-module income "summary screen", designed to allow respondents to review and edit their reports of income. The aim of this summary screen is improving income data quality by reducing the level of item missing data (which are typically high in income data), reducing the number of outliers and the deviation of outliers from the mean and add the missing sources of income that the respondent may have failed to report. We also tested

various in-interview consistency checks and item non-response follow-ups, and the inclusion of additional response categories for benefit and income sources.

Research questions

- Are respondents invited to complete the income summary screen willing to review and revise their reports? If so, who are the respondents willing to do so?
- How does the income distribution change? And which sources of income are revised and in which direction?
- Why is the summary screen effective? Does it lead to revision of amounts, to adding sources and/or deleting sources?
- Can motivational statements decrease item non-response in income in web surveys? Can the use of closed responses decrease item non-response in income? Do consistency checks decrease misreporting?

Experimental design: the "Income summary screen"

At the end of the income module, a random half of the understanding society sample members (N=1,096) were presented with the editable income summary screen.

The summary screen presented respondents with all of their reported income sources. Specifically, information included in the summary screen were: employment income (from main job, second job, self-employment), benefits and pension income (e.g. State pension, Employer pension, Working Tax Credit, Child Benefit, Child Tax Credit, Disability Living Allowance, any other benefit), and any other income. Respondents could correct (upwards/downward) the amounts and add or delete sources.

Respondents with inapplicable income sources or missing data in all questions on income were not presented with the summary screen. The summary screen was not presented to respondents conducting their interview by telephone; respondents being interviewed by web completed their summary screen in web self completion, while respondents interviewed by face-to-face had the laptop handed over to them by the interviewer.

The amounts reported in the summary screen are net amounts and they are converted in monthly equivalent when the respondent has provided in the income module figures for a timeframe of four weeks. All other in-interview consistency checks and innovations (see section below) fed into the end-of module summary screen.

Experimental design: the in-interview consistency checks and other innovations

In the income module a series of consistency checks and other innovation were implemented; specifically, these are:

- Item non response follow-ups (web)

  In the web interview, respondents attempting to skip a survey question on income were presented, before the refused/don't know category is displayed, with a motivational statement, informing them about the importance of the figure for the quality of the study.

- Item non response follow-ups (closed response categories)

  In face-to-face interviews, if self-employed respondents refused to provide an answer in the survey question on net profit in last yearly account – or if he/she reported "don't know" – the interviewer shows the respondent a showcard with ranges of income (i.e. less than £250, £250 to £499, …, £10,000 or more) and asks him/her to choose an answer category. In web interviews, the closed response categories are displayed in the screen.

- Benefit consistencies checks

  A follow-up question was added for sample members who reporting receiving a benefit when other data suggested that the respondent is not eligible for receiving such benefit. Similarly, respondents not reporting receiving a benefit were presented with a follow-up question when other survey data suggesting the respondent is eligible for that benefit. For example, respondents reporting as a source of income a state retirement pension, were followed up with a "consistency check" question if according to the reported age they were not entitled to receive a state retirement pension.

- Net-gross consistencies checks

  A follow-up question was added for respondents reporting net income larger or equal to their gross income.

- Income and benefit - new response categories

The list of income and benefit sources was expanded to include categories which, in previous waves, were not explicitly mentioned. In previous waves respondents were expected to include income from these categories into the residual "other income" item. Examples of these categories are: royalties (e.g. land, books or performances), annuity (includes home income plan or equity release), and occupational pensions from an overseas government/company.

- Net-gross checks

  Levels of net earnings were asked for categories of earnings which in previous waves were only collected as gross earnings (e.g. net earnings from second jobs last month, tax paid on interest/dividends).

Results

A substantial 11.5 percent of respondents confirmed that the income totals presented to them in the summary screen were wrong. Moreover, respondents were willing to correct their reports when asked to do so (hereafter "correctors"). Respondents were willing to fill in missing amounts, revise incorrect amounts and also to add income sources. Compared to the "confirmers", the correctors tended to be: slightly older (53.7 vs. 49.6), more likely to be retired (37.7 vs. 27.1 percent), have more sources of income (2.8 vs 2.1) and be poorer as defined by their initial income reports (15 ppt more likely to be in the 2nd income decile).

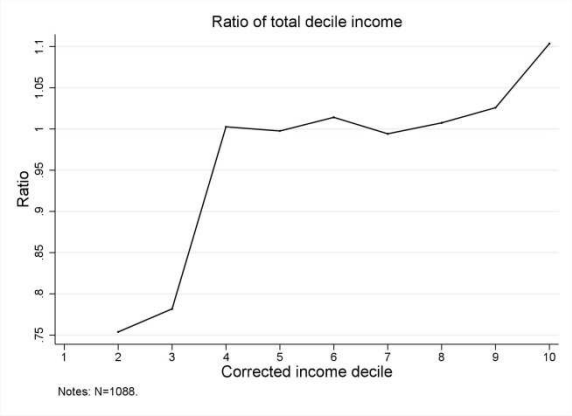
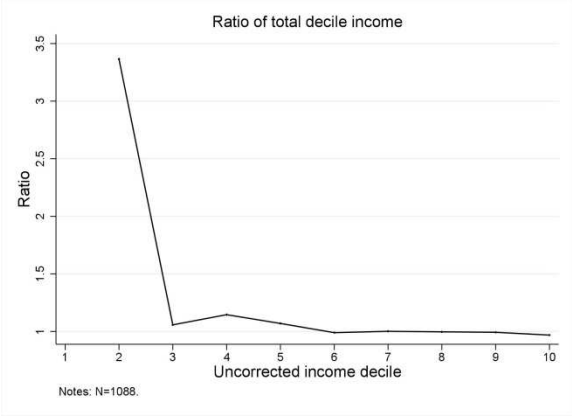
Of the correctors, 58 percent revised their total income downwards, while 41 percent revised in the opposite direction. The increases tended to be larger than the decreases, and so reported income increased on average by £257.96 per month or 17 percent. Given what we know about income being under-reported in household surveys, and the fact that item missingness fell, this suggests that the income summary screen was effective at improving data quality.

The biggest share of corrections occurred in benefits and unearned income (75.7 percent), followed by employee earnings (24.4 percent), second job earnings (9.6) and self-employment income (4.4). The magnitude of the corrections follows a different pattern: increases in earnings explain the majority of the increase in reported income (93.9 percent of the total increase in reported income), followed by self-employment income 10.8 percent, second job 4.8 percent, while benefits and unearned income were corrected down on average and represent -9.5 percent of the total change.

Figure 1 shows how the summary screen changes what we know about the individual total income distribution. Panel A shows for each decile of original income, the mean growth in income due to the income revisions at the summary screen. We see that right across the distribution, mean income growth is positive but the largest (proportional) increases occur at the bottom of the initial reported distribution. For example, those in the third decile of original income saw their incomes increase by on average 5.7 percent, while those in the 2nd decile saw it increase by a substantial 337 percent.

Panel B shows where the correctors come from in the true or corrected income distribution by plotting income growth for each corrected income decile. It is individuals with the highest incomes who originally under-estimated their income, whilst individuals in the bottom half of the income distribution originally over-estimated their income. For example, individuals in the top income decile increased their original report by 10 percent and individuals in the second decile reduced their original report by 25 percent.

Figure 1: correction in reported income by (corrected and uncorrected) income decile

Panel A	Panel B
	

Note: ratio = corrected income/uncorrected income

In addition to the income summary screen, the IP9 income module included other innovations and consistency checks to improve income data quality. Motivational statements decreased substantially item non response in the web component of the mixed mode survey.

Also, the use of closed response categories in the face to face survey (though banded showcards) and in the web survey, decreased non-response further.

Consistency checks on net and gross income identified that net income is greater or equal to gross income for 16% of respondents reporting both gross and net income; most respondents confirmed that this information is correct; among those not confirming these figures, the majority agreed to revise their estimate.

We asked a follow-up question when respondents didn't report benefits that we expected them to receive (based on other information) or, the other way around, they reported benefits that we didn't expect them to receive. These follow-up questions are applicable to a small fraction of respondents; nevertheless, in most cases the respondent reiterates that the information is correct; only for child benefit we have 30% that report the information collected is wrong.

Conclusions

Put together, we conclude that misreporting of income is not fixed feature of data collection and an income summary screen can be very effective at improving data quality.

Furthermore, some additional in-interview innovation and consistency checks are successful in improving income data quality.

(3) Masking opposition to immigration: an experimental approach to understand the dynamics of social desirability bias (Mathew J. Creighton, and Amaney Jamal)

There is a growing body of literature showing that intolerance is masked from direct questioning. Little is known about the dynamics of masking across time. We use the Item Count Technique (ICT), also known as the list experiment, to manipulate the level of anonymity offered to respondents. As a comparison, we pose a direct question to three separate control groups that corresponds to the questions measured with the list experiment. In addition, these control groups received a list question that did not include the focal item of interest, which provides a baseline distribution for comparison. 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 express opposition to immigrants from the same country-type origins. This defined the covert opposition these same groups confront. Third, we compare the over and covert proportions to ascertain the proportion of the population that masks their opposition. This captures the level of social desirability bias (SDB). The results that follow are based on identical questions as those embedded in IP8 and are intended to assess post-referendum stability in masking of attitudes toward Eastern European, Caribbean, and Muslim immigrants to the UK.

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):

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

where S 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):

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

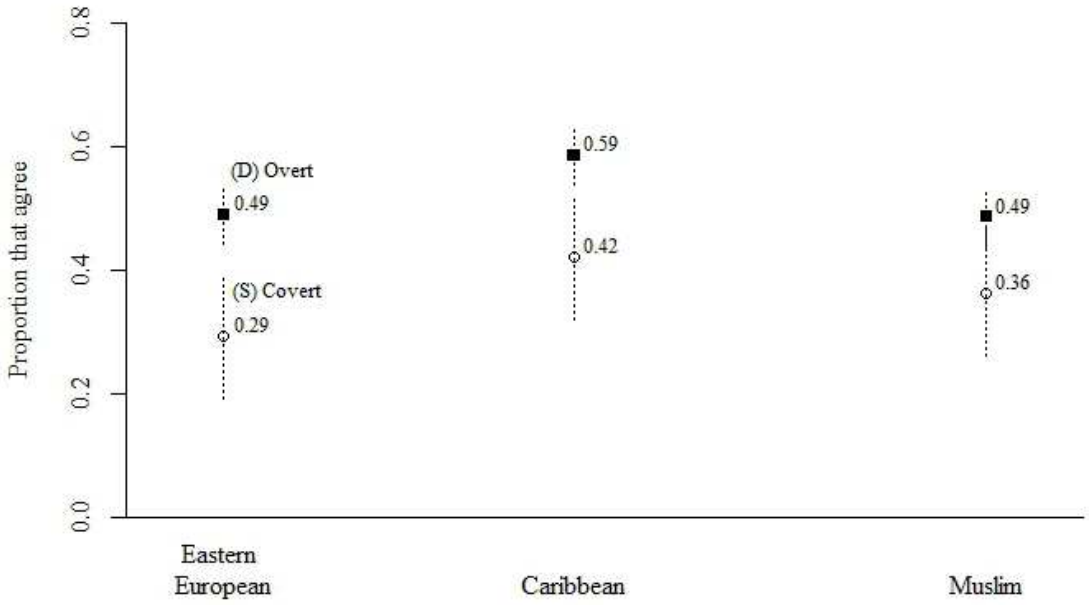
where B is direct measure of SDB that, when converted to a percentage scale, is typically interpreted as the number of percentage points difference between the explicit, derived from the control sample, and the implicit estimate (S), derived from equation (1).

Preliminary Results:

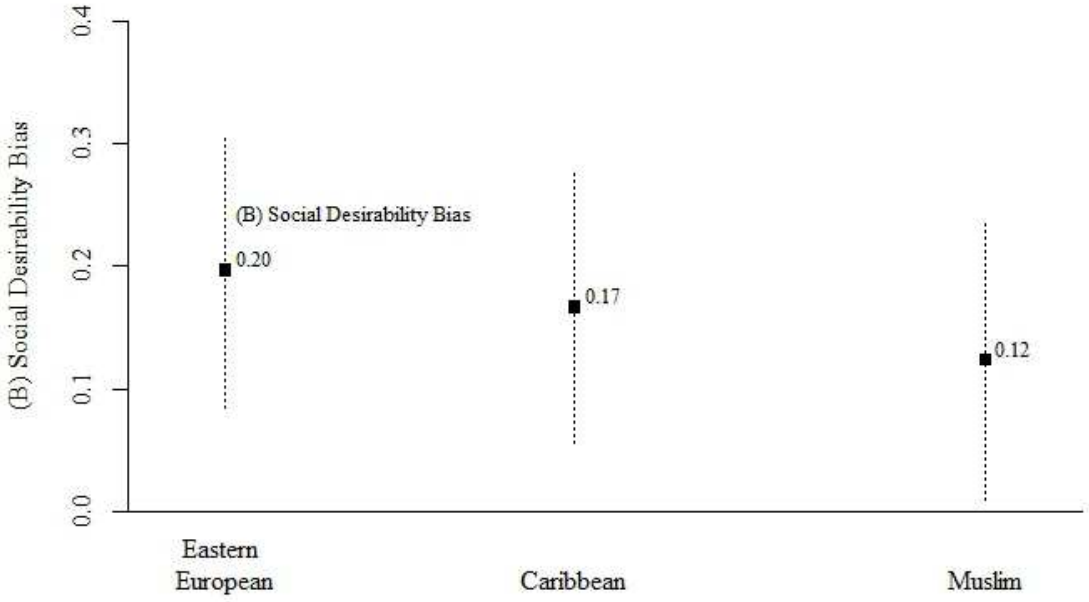
Plot 1 reports the estimated proportion in favor of allowing immigrants of each of the three country-origin types assessed by the experiment. The direct measure ($\bar{X}_D$), which is higher than the covert estimate in each case, is directly estimated. The list proportion (S) is estimated using equation (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).

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



Plot 2: Social Desirability Bias



(4) A comparison of self-reported sexual identity using direct and indirect questioning (Alessandra Gaia)

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 a 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), I 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 sample who have the sensitive characteristic (Tourangeau & Yan 2007:872).

The wording of the ICT questions is detailed in Table 8.

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 HIS uses an interviewer administered approach with a show card, if face-to-face, or no show card when over the telephone[13].

[13] 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

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 IP9 mixed-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 less than 0.4% of respondents skipping the survey question. Refusal was also not frequent, ranging from 2.3% (n=24) of respondents to 0.5% (n=5) respondents; and don't know answers were rare, ranging from to 0.6% (n=6) to 0.2% (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 (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").

---

in telephone interviews. We refer to this protocol as "UKHLS", as this is the protocol currently adopted 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 "don'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.

Unexpectedly, but consistently with IP8, in the fields of attraction (lists A and B) and behaviour (lists C and D), the relative majority (over 29%) 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.

Table 8. 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*	*4*	*0.39*
Refusal		*13*	*1.23*	*18*	*1.74*
don't know		*6*	*0.57*	*2*	*0.19*
• is the same sex as me	None of the statements are true	311	29.31	317	30.66
• has a disability	One of the statements are true	238	22.43	254	24.56
• is fit and muscular	Two of the statements are true	253	23.85	240	23.21
• grew up with me in my local area	Three of the statements are true	185	17.44	126	12.19
• is ten or more years older than me	Four of the statements are true	54	5.09	62	6
How many statements are true for you?	Five statements are true	N.A.	N.A.	11	1.06
		LIST B		LIST B+S	
I have at least once been sexually **attracted** to someone who …		Obs	%	Obs	%
missing		*4*	*0.39*	*1*	*0.09*
refusal		*18*	*1.74*	*11*	*1.04*
don't know		*6*	*0.58*	*6*	*0.57*
• is the same sex as me	None of the statements are true	355	34.33	344	32.42
• wears the latest trends and fashions	One of the statements are true	198	19.15	166	15.65
• has a tattoo or body piercing	Two of the statements are true	171	16.54	177	16.68
• is of a different ethnicity to me	Three of the statements are true	116	11.22	143	13.48
• is from a different class background to me	Four of the statements are true	166	16.05	165	15.55
How many statements are true for you?	Five statements are true	N.A.	N.A.	48	4.52

Table8. 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*	*4*	*0.39*
refusal		*18*	*1.7*	*24*	*2.32*
don't know		*4*	*0.38*	*5*	*0.48*
• is the same sex as me	None of the statements are true	367	34.59	362	35.01
• has a disability	One of the statements are true	330	31.1	292	28.24
• is fit and muscular	Two of the statements are true	221	20.83	207	20.02
• grew up with me in my local area	Three of the statements are true	102	9.61	100	9.67
• is ten or more years older than me	Four of the statements are true	18	1.7	34	3.29
How many statements are true for you?	Five statements are true	N.A.	N.A.	6	0.58
		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		*4*	*0.39*	*1*	*0.09*
refusal		*24*	*2.32*	*19*	*1.79*
don't know		*3*	*0.29*	*5*	*0.47*
• is the same sex as me	None of the statements are true	436	42.17	436	41.09
• wears the latest trends and fashions	One of the statements are true	218	21.08	203	19.13
• has a tattoo or body piercing	Two of the statements are true	139	13.44	158	14.89
• is of a different ethnicity to me	Three of the statements are true	102	9.86	105	9.9
• is from a different class background to me	Four of the statements are true	108	10.44	107	10.08
How many statements are true for you?	Five statements are true	N.A.	N.A.	27	2.54
IDENTITY		LIST E		LIST E+S	
I would describe myself as being …		Obs	%	Obs	%
missing		*1*	*0.09*	*4*	*0.39*
refusal		*5*	*0.47*	*8*	*0.77*
don't know		*2*	*0.19*	*3*	*0.29*
• gay, lesbian or bisexual	None of the statements are true	38	3.58	55	5.32
• stylish and fashionable	One of the statements are true	232	21.87	256	24.76
• disabled	Two of the statements are true	538	50.71	450	43.52
• patient	Three of the statements are true	228	21.49	242	23.4
• British	Four of the statements are true	17	1.6	14	1.35
How many statements are true for you?	Five statements are true	N.A.	N.A.	2	0.19
		LIST F		LIST F+S	

I would describe myself as being …		Obs	%	Obs	%
missing		*4*	*0.39*	*1*	*0.09*
refusal		*6*	*0.58*	*6*	*0.57*
don't know		*5*	*0.48*	*3*	*0.28*
• gay, lesbian or bisexual	None of the statements are true	33	3.19	34	3.2
• healthy	One of the statements are true	154	14.89	166	15.65
• tolerant	Two of the statements are true	275	26.6	276	26.01
• European	Three of the statements are true	373	36.07	355	33.46
• working class	Four of the statements are true	184	17.79	206	19.42
How many statements are true for you?	Five statements are true	N.A.	N.A.	14	1.32

As for IP8, the evidence on the "ceiling effect" is mixed; while lists A (attraction), C (behaviour) and E (identity) resulted 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 that all four behaviours ranges between 10.4 and 17.8%, 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 A+S", and "List E+S"), compared with the average in the list without the sensitive items – "List A", "List E"(see table 9).

Vice versa, and consistently with our expectations, in the "behavioural" questions we observed an 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 9.1% (as opposed to 9.9% at wave 8), see table 9.

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 showed in table 10 there are no statistically significant difference across the two protocols.

Table 9. The estimates from the Item Count Technique

Average "List A"	Average "List A+S"	Average "List A+S" – Average "List A"
1.40	1.46	-0.05
Average "List B"	Average "List B+S"	Average "List B+S" – Average "List B"
1.77	1.54	0.23
Estimated prevalence of homosexual/bisexual attraction: N.A.		
Average "List C"	Average "List C+S"	Average "List C" – Average "List C+S"
1.17	1.11	0.06
Average "List D"	Average "List D+S"	Average "List D" – Average "List D+S"
1.35	1.23	0.12
Estimated prevalence of homosexual/bisexual experience: 9.1%		
Average "List E"	Average "List E+S"	Average "List E+S" – Average "List E"
1.90	1.35	0.56
Average "List F"	Average "List F+S"	Average "List F+S" – Average "List F"
2.43	2.56	-0.14
Estimated prevalence of homosexual/bisexual identity: N.A.		

Table 10. A comparison of the UKHLS and IHS protocols

	UKHLS				IHS			
	Obs.	%	95% C.I.		Obs.	%	95% C.I.	
heterosexual/straight	1459	91.02	89.11	91.02	500	91.91	86.61	95.23
gay/lesbian	26	1.62	1.03	2.54	9	1.65	0.76	3.57
bisexual	31	1.93	1.34	2.79	9	1.65	0.81	3.37
other	16	0.01	0.61	1.63	2	0.37	<0.01	1.51
prefer not to say	52	3.24	2.28	4.61	N.A.	N.A	N.A	N.A
don't know	2	0.12	<0.01	0.90	12	2.21	0.79	6.00
refusal	12	0.75	0.42	1.33	12	2.21	0.84	5.64
missing	5	0.31	0.16	0.62				
total	1464				544			

Note: the category "prefer not to say" is not displayed in the IHS version a 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).

(5) Separating systematic measurement error components using multi-trait multi-error (MTME) in longitudinal studies (Alexandru Cernat and 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 (Al Baghal *et al*. 2016). Below we will present the first results from wave 9 of the Innovation Panel.

We used 6 items that measure attitudes towards immigrants.

Table 11. Traits and Social Desirability Direction for MTMM Experiment

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

There are 8 different wordings of each item, corresponding to combinations of three factors: the higher- or lower-end being the socially desirable direction, the number of scale points, and whether an agree-disagree or disagree-agree question are used. These lead to 8 wordings W1-W8; an example formulation for trait one is given in the last column of Table 12.

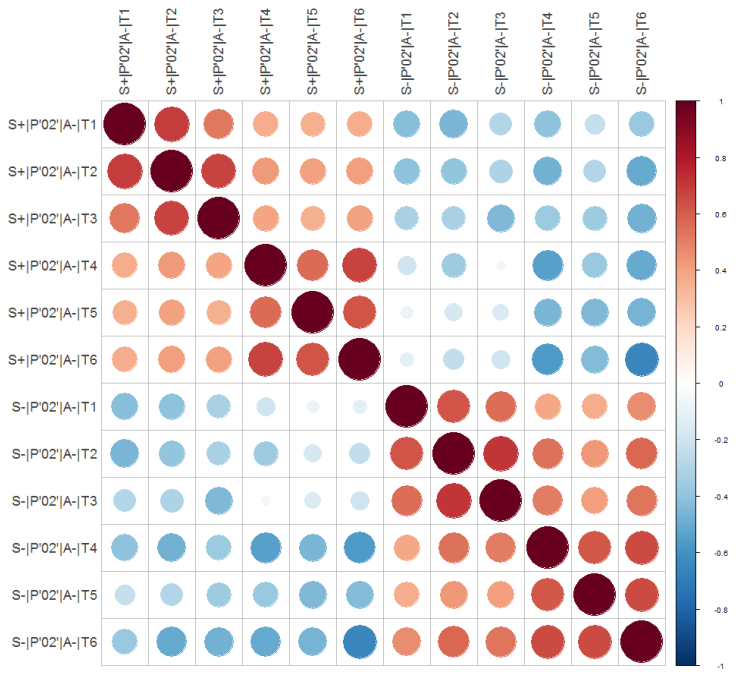
Table 12. Item Wordings for MTMM Experiment

Wording number	Social desirability	Number of scale points	Agree or Disagree	Required direction	Item formulation (using trait 1 as an example)
W1	Higher	2	AD	Negative	The UK should allow **fewer** people of the same race or ethnic group as most British people to come and live here
W2	Lower	2	AD	Positive	The UK should allow **more** people of the same race or ethnic group as most British people to come and live here
W3	Higher	11	AD	Negative	The UK should allow **fewer** people of the same race or ethnic group as most British people to come and live here
W4	Lower	11	AD	Positive	The UK should allow **more** people of the same race or ethnic group as most British people to come and live here
W5	Higher	2	DA	Positive	The UK should allow **more** people of the same race or ethnic group as most British people to come and live here
W6	Lower	2	DA	Negative	The UK should allow **fewer** people of the same race or ethnic group as most British people to come and live here
W7	Higher	11	DA	Positive	The UK should allow **more** people of the same race or ethnic group as most British people to come and live here
W8	Lower	11	DA	Negative	The UK should allow **fewer** people of the same race or ethnic group as most British people to come and live here

Initial results

The correlation plot bellow 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.

Figure 2.Correlation matrix with 2 wordings of the 6 items.



Note: "S+" indicates higher Social Desirability and "S-" lower. Larger circles and darker colours indicate stronger relationships.

Here we use two different wording of the survey questions in order to manipulate social desirability ("S+" *vs.* "S-"). 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 ("S+"). 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 ("S-"). It is expected 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 their answers are given using a 2 point scale (P'02') using the Agree-Disagree order of the categories ("A-") for all 6 items (T1-T6).

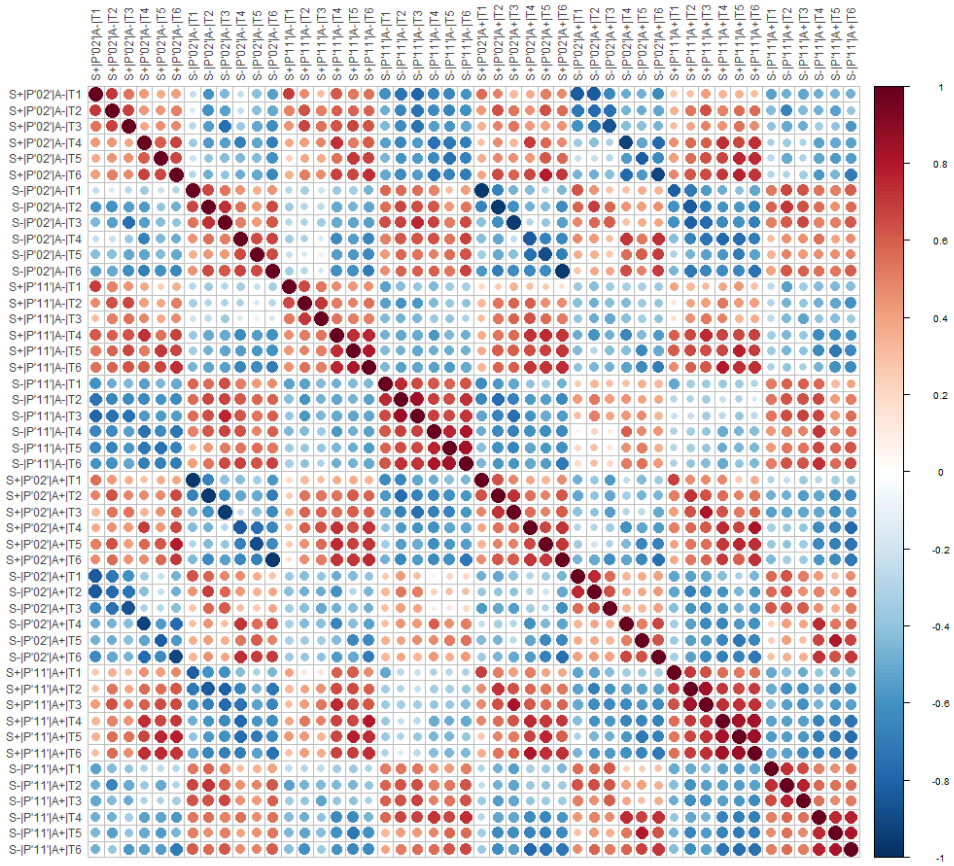
In the top left corner we can see how strong are the relationships between our 6 items. For example, we can see that the relationship between the first 3 items is stronger (larger circles and more intense red colour) than the one between the variables 4-6. This is most likely because they come originally 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 we see 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 has 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).

Figure 3.Correlation matrix with 8 different wordings of the 6 items.

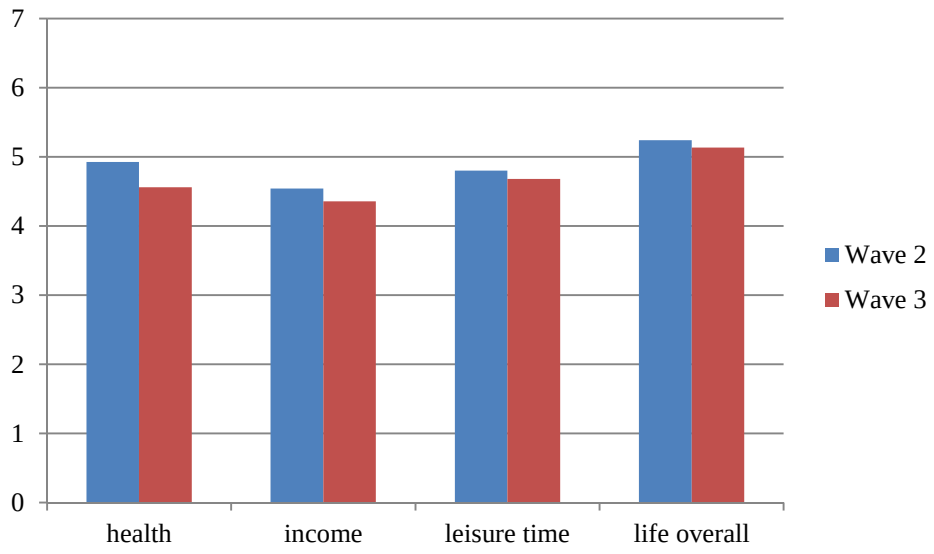


Note: T = traits, S = social desirability, P = response scale, A = acquiescence

(6) IP9 experiment on visual presentation of satisfaction scales (Jonathan Burton)

On the *Understanding Society* survey, participants are asked for their feelings of satisfaction in different domains; their health, the income of their household, the amount of leisure time they have, and their life overall. These questions are asked in a self-completion mode. When the self-completion instrument switched from a paper questionnaire at Waves 1 and 2, to a computer-assisted self-interviewing (CASI) instrument at Wave 3, a fall in the mean level of 'satisfaction' was observed. Graph 1, below, shows the average satisfaction score at Wave 2 and Wave 3, this is a balanced panel and Wave 3 longitudinal weights are used. The scale ranges from 1 "Completely dissatisfied" to 7 "Completely satisfied". The differences are significant with no overlap in the 95% confidence intervals.

Graph 1: Mean level of satisfaction (1-7), Waves 2 and 3

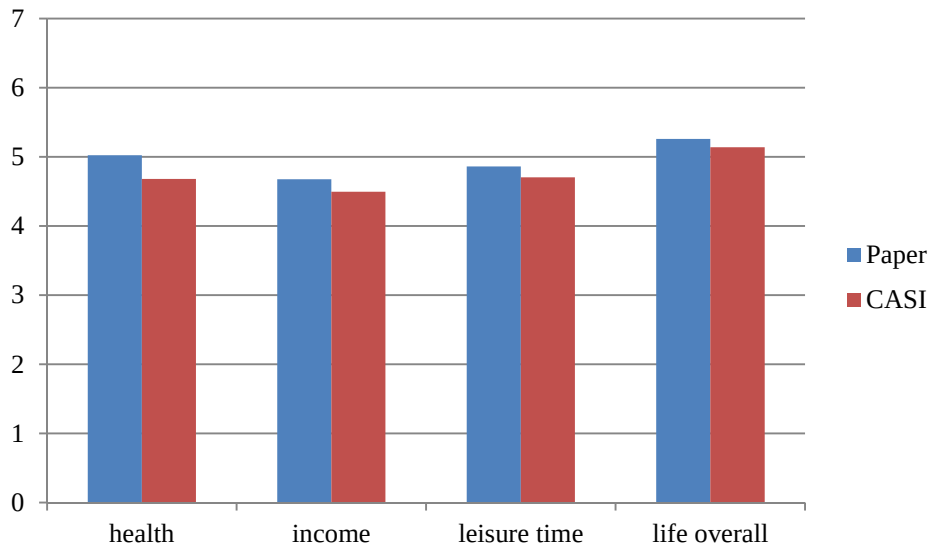


In the fourth wave of the Innovation Panel (IP4) an experiment on the mode of the self-completion instrument was conducted, with adults in a randomly allocated half of households receiving a paper questionnaire, and the adults in the other half receiving the questions in CASI. Details of this experiment are available in Burton *et al* (2012)[14]. Participants who completed the satisfaction questions on paper had a higher mean satisfaction than those who completed using CASI. This echoes the finding from the comparison of Waves 2 and 3 of the main-stage.

It should be noted that it was not only the mode of the self-completion instrument that was different at Wave 3 (compared to Wave 2) and in CASI compared to paper at IP4. The limitations of the CASI software at the time meant that the questions and responses were presented differently in CASI than on paper. On paper, they are presented as a grid, with each domain as a row and the response categories running horizontally. In CASI, the questions were presented one at a time, with the response options running vertically.

[14] https://www.understandingsociety.ac.uk/research/publications/working-paper/understanding-society/2012-06

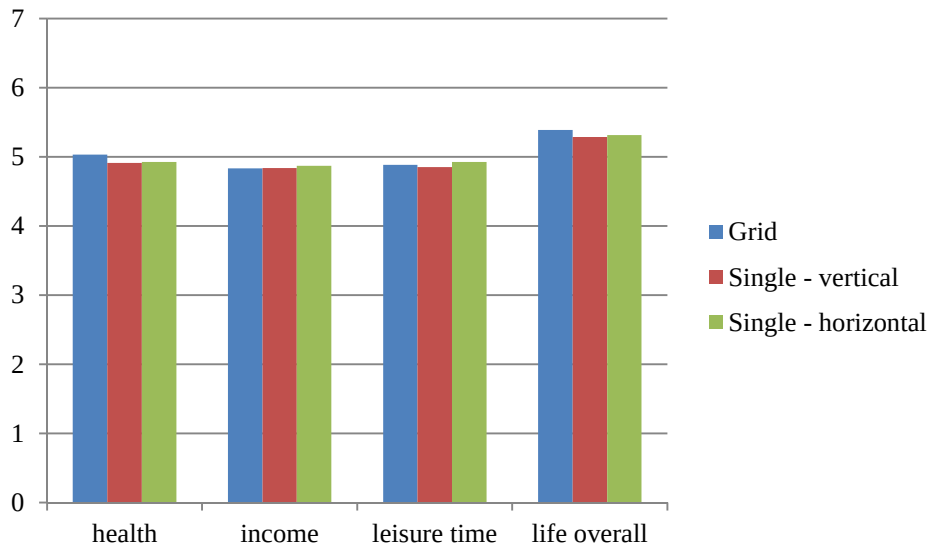
Graph 2: Mean level of satisfaction (1-7) by mode of self-completion, IP4



The present IP9 experiment uses three different presentations of question and response options on the standard set of satisfaction questions. These were all administered as self-completion, either using CASI (for those interviewed face-to-face), or online. One version had all satisfaction items presented on one screen, with response options across the top and the items on the left in a grid format (similar to the paper instrument at Waves 1 and 2, and the paper treatment group at IP4). The second version had each question presented on a separate screen, with response options vertically aligned (similar to the Wave 3 main-stage and the IP4 CASI treatment group). The final version also had each question presented on a separate screen, but response options will be presented horizontally (combining the question-by-question design of version 2, with the alignment of version 1). Screenshots of these layouts are available in the appendix.

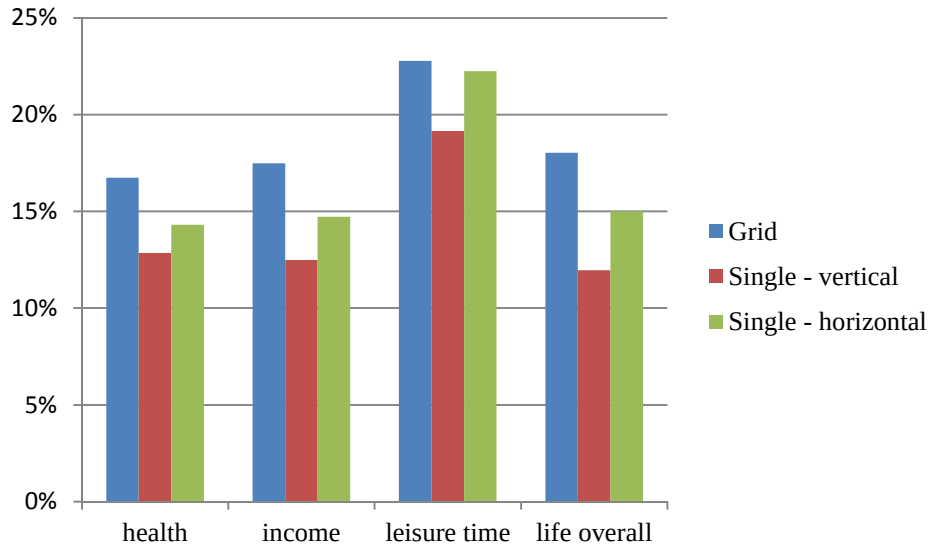
The goal of the experiment was to identify if and how the presentation of response options in self-completion formats can impact estimates of satisfaction. However, we find that there is no statistical difference between the means across the treatment groups for any of the domains of satisfaction. Further analyses of treatment group and mode of interview (face-to-face or web) and also of mean satisfaction by mode of interview also show no statistically significant differences.

Graph 2: Mean level of satisfaction (1-7) at IP9 by treatment group.



There was some evidence that horizontal scales led to more extreme reporting (either completely satisfied or completely dissatisfied) compared to vertical reporting. There is a statistically significantly ($p<0.05$) higher level of extreme reporting in the grid version than the vertical version for health, income, and life overall. The single-horizontal version shows higher levels of extreme reporting than the vertical scale, but this is not statistically significant.

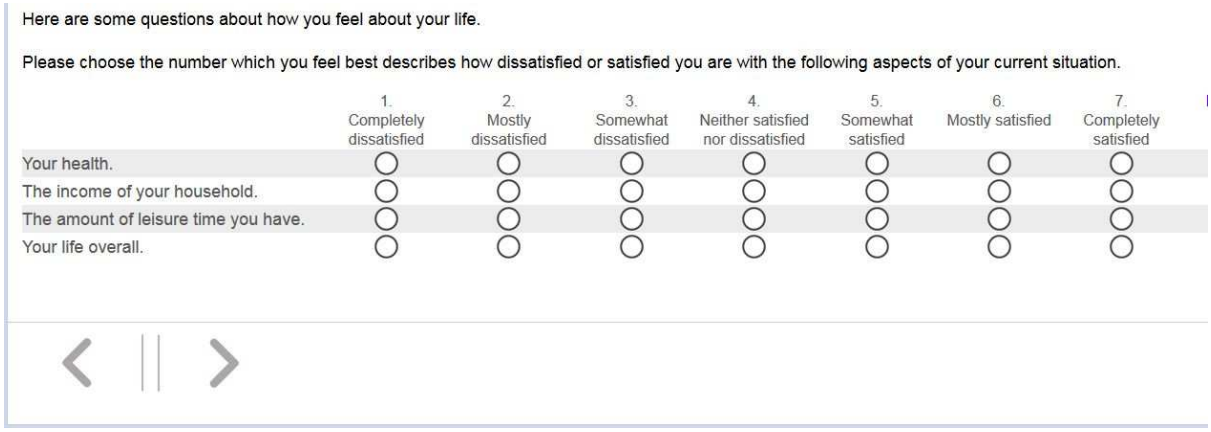
Graph 3: Percentage of cases with 'extreme reporting' at IP9 by treatment group



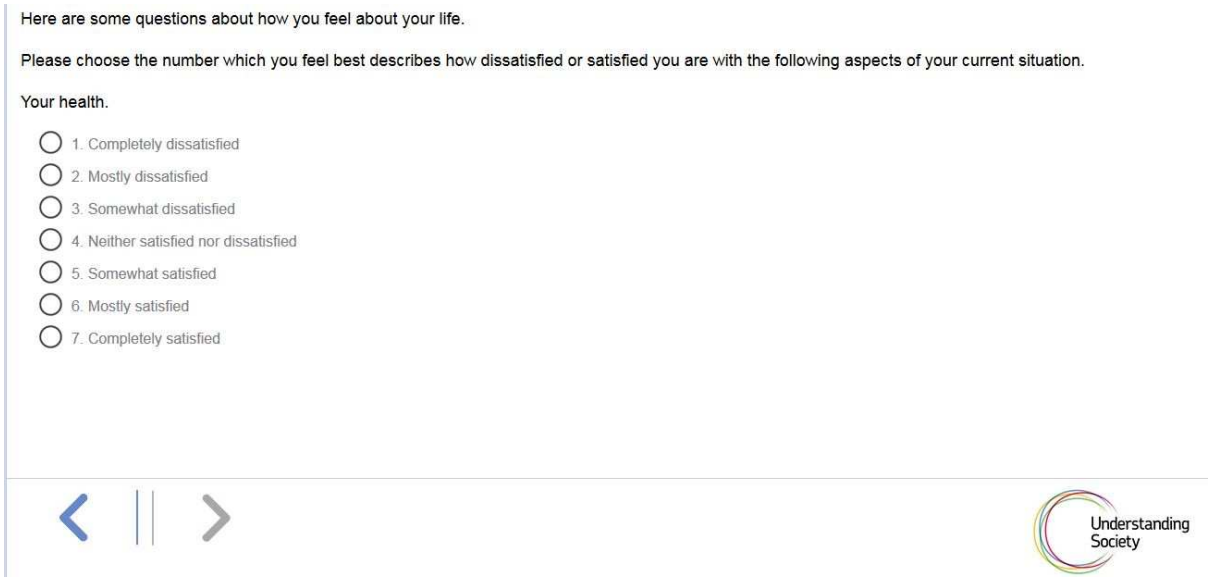
Note: "extreme reporting" is designated as responding "completely" satisfied or dissatisfied.

Appendix: Screenshots of satisfaction questions

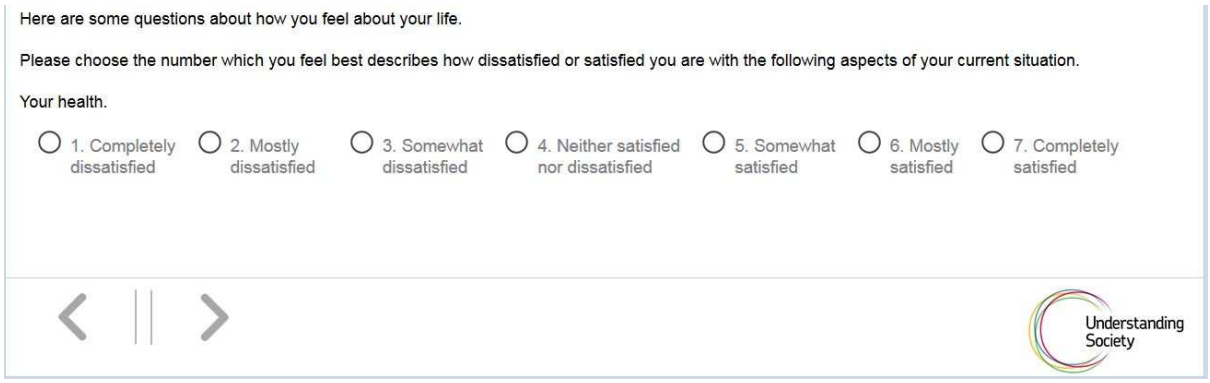
Group 1 – Grid.



Group 2 – Single screen, vertical responses



Group 3 – single screen, horizontal responses



(7) What do the general population regard as "successful ageing"? (Elise Whitley, Michaela Benzeval and Frank Popham)

Industrialised populations are ageing (Christensen, Doblhammer, Rau, & Vaupel, 2009) and medical advances mean that increasing numbers of, particularly older, individuals are living with disease and disability. There is therefore substantial interest in what constitutes "successful ageing" (SA) (A. Bowling, 2007; Ann Bowling & Dieppe, 2005; Katz & Calasanti, 2014; Kivimaki & Ferrie, 2011; Martin *et al*., 2014). SA is an important goal for health and economic policies (Bloom *et al*.; CEC, 2009; UN, 2002), and effective measurement is vital for understanding the challenges, informing potential interventions, and monitoring progress towards its achievement. Many definitions of SA have been proposed by researchers (Cosco, Prina, Perales, Stephan, & Brayne, 2013; Depp & Jeste, 2006) but no consensus has been reached and evidence suggests that many older people who consider themselves to be ageing successfully do not meet researcher-defined SA criteria (McLaughlin, Jette, & Connell, 2012; Montross *et al*., 2006; Strawbridge, Wallhagen, & Cohen, 2002; von Faber *et al*., 2001; Young, Frick, & Phelan, 2009). The SA model most widely used in the literature was proposed by Rowe and Kahn (Rowe & Kahn, 1997), and incorporates six dimensions: (i) avoidance of disease (ii) avoidance of disability; (iii) maintenance of good physical function; (iv) maintenance of good cognitive function; (v) good interpersonal social engagement (contacts and transactions with others); and (vi) good productive engagement (engagement in activities of value to society such as working or volunteering). However, the extent to which this model reflects perceptions of SA in the general population is still largely unknown. Existing work on people's views on SA has been

primarily qualitative, with individuals giving responses to open ended questions. However, it is well recognised that responses to such questions are influenced by individuals' own experiences and circumstances. An alternative approach (Wallander, 2009) is to use standardised vignettes (descriptions of a fictitious third party) in which factors are systematically varied. In our experiment, rather than rating their own ageing experience, respondents were asked to rate the SA of a third party described in terms of the Rowe-Kahn dimensions. This approach has not been used previously in this context and provides a unique empirical assessment of the functionality of the Rowe-Kahn model.

Methods

Respondents were presented with three distinct vignettes, each describing a 75 year old person with favourable or unfavourable outcomes in each of the six dimensions of the Rowe-Kahn SA model. After each vignette, respondents were asked "How successfully is [name] ageing?" and gave an SA score on a scale of 0 (not successfully) to 10 (very successfully). Descriptions in the vignettes aimed to be neutral, incorporate conditions and circumstances likely to be well recognised in the context of SA, and open to interpretation in terms of their likely impact on SA. Favourable/unfavourable outcomes in each dimension were:

- *Chronic disease:* no long-term illness / diabetes;
- *Disability:* no difficulties / difficulties climbing stairs;
- *Physical functioning:* opens food packages easily / struggles to open food packaging;
- *Cognitive functioning:* no problems / problems remembering;
- *Interpersonal engagement:* regularly / rarely sees friends and family;
- *Productive engagement:* often volunteers / doesn't volunteer.

An example vignette is as follows:

George is 75 and has diabetes.

He has difficulties climbing stairs, no problems remembering, and opens food packages easily

He rarely sees family and friends and often volunteers.

How successfully is George ageing?

Vignettes were randomly allocated at the individual level using a factorial design to ensure that all SA dimensions were equally represented across all respondent characteristics (age, gender) and data collection mode – web (CAWI) or computer-assisted personal interviewing (CAPI) interviews. The relative weightings given to each SA dimension were explored using standard methods (Atzmueller & Steiner, 2010), comparing SA scores for vignettes in which the dimension was favourable with scores for vignettes in which the dimension was unfavourable.

Results

Of 2,174 respondents included in IP9, 2,143 (99%) took part in either CAWI or CAPI interviews and, of these, 2,010 (94%) agreed to take part in the self-completion module, which included the vignettes. Characteristics of respondents presented with vignettes with favourable/unfavourable SA dimensions are presented in Table 13. Approximately equal numbers of vignettes had favourable/unfavourable dimensions and there were no marked differences in the gender or age of respondents receiving them or the mode of delivery, confirming that the randomisation resulted in a balanced design. Of those presented with the vignettes, 1,986 (99%) responded to all three and just 24 (1%) to two or fewer, suggesting good acceptability and engagement with the exercise. In total, SA scores were available for 5,967 vignettes and the mean (standard deviation) SA score across all vignettes was 6.2 (2.3). The range of SA scores given by each respondent, i.e. the difference between the lowest and highest score given to the three vignettes, varied from 0 to 10 with a mean value of 2.8, demonstrating that respondents did not simply allocate the same SA score to all vignettes.

Initial results (Table 14) indicate that respondents gave consistently higher SA scores to vignettes in which SA dimensions were favourable. However, these differences were not consistent across all dimensions. The largest differences were observed in vignettes comparing favourable with unfavourable cognitive functioning and disability and the smallest differences were observed in vignettes comparing favourable with unfavourable productive engagement and chronic disease. Further planned analyses will consider how respondents' views of SA vary according to gender (their own and that of the person described in the vignette), age, interview mode, and circumstances, such as ill health and satisfaction with aspects of life. We will also explore the potential of a weighted Rowe-Kahn SA measure, based on population-based weightings derived from this experiment.

Table 13: Characteristics of respondents presented with favourable/unfavourable dimensions

SA dimension	Comparing favourable / unfavourable			
	N vignettes	% male	% aged <50	% CAPI
Chronic disease	3,083 / 2,977	44.6 / 45.6	47.6 / 48.2	44.3 / 44.4
Disability	3,056 / 2,974	45.8 / 44.4	47.4 / 48.4	45.8 / 42.9
Physical functioning	3,025 / 3,005	45.3 / 44.8	47.7 / 48.0	43.8 / 44.9
Cognitive functioning	3,059 / 2,971	44.1 / 46.1	48.1 / 47.6	44.2 / 44.5
Interpersonal engagement	3,000 / 3,030	44.8 / 45.4	47.8 / 47.9	43.8 / 45.0
Productive engagement	3,035 / 2,995	44.6 / 45.6	47.9 / 47.8	43.7 / 45.1

Table 14: Difference (95% CI) in SA score comparing favourable with unfavourable SA dimensions

SA Dimension	Mean (SE) SA score for negative vignettes	Mean (SE) SA score for positive vignettes	Difference (95% CI) in SA score
Chronic disease	5.83 (0.04)	6.56 (0.04)	0.72 (0.61, 0.84)
Disability	5.58 (0.04)	6.81 (0.04)	1.23 (1.12, 1.35)
Physical functioning	5.79 (0.04)	6.61 (0.04)	0.83 (0.71, 0.94)
Cognitive functioning	5.58 (0.04)	6.81 (0.04)	1.24 (1.12, 1.35)
Interpersonal engagement	5.68 (0.04)	6.73 (0.04)	1.06 (0.94, 1.17)
Productive engagement	5.90 (0.04)	6.51 (0.04)	0.61 (0.50, 0.73)

(8) Targeted weekday of the week to send email invitation (Annamaria Bianchi)

Survey research literature has long recognized that many aspects of a way a survey is carried out can affect response rates. One of those aspects is timing of contact. Most research on optimal contact scheduling has been carried out in the context of telephone surveys or face-to-face surveys (Durrant *et al.*, 2011; Wagner, 2013), highlighting in general the importance of timing of calls.

At IP9, an experiment was implemented to test the effectiveness of targeted weekday of invitation emails. A random half of the web sample received the email invitation based on standard procedure (control group). This is the approach that has been taken at previous waves of the IP. Sample members in the other half of the IP9 web sample were sent the invitation e-mail depending on which day they preferably responded to the questionnaire in previous waves of the panel. Taking into account that household members are requested to answer a household grid, a household questionnaire, and individual questionnaires, and that for members of the households receiving invitations on different days could be confusing, the

experiment was set up as follows. The initial invitation was sent on the same day to all household members, identifying a preferred day for the household based on the day the household questionnaire was completed in past waves. Subsequent personal remainders were sent according to past responding days of each individual.

The aim of the experiment was to assess the overall effect of the targeting strategy on web response rate and response speed. Initial estimates of the effects on web response rates are presented here. Tables 15 and 16 report results for household questionnaire web response and individual web response, respectively. Results are reported for the overall sample and for the original and IP4 refreshment sample separately.

Across the issued adult sample as a whole, the household questionnaire response rate at IP9 was 55.3% for the control group and 58.9% for the targeted group (P=0.39; N=580). Looking at individual response, it is 67.0% for the control group and 68.1% for the targeted group (P=0.72; N=944). There is therefore no evidence of an overall effect on response rate by web.

Splitting the sample into groups, defined by the number of waves sample members are in the panel, no significant differences could be found in the original sample (for whom this is the ninth wave). As for those who entered the panel at wave 4 (IP4 refreshment sample), the household questionnaire response rate was 16 percentage points higher for the targeted group with respect to the control group (72.8% vs 56.8%, P=0.03; N=173). No significant differences could be detected at the individual level.

Table 15. Household questionnaire response rates between standard and targeted timing protocol, overall and by sample.

	Targeted	Standard	P	N
All	58.9	55.3	0.39	580
Original Sample	53.1	54.6	0.76	407
IP4 Refreshment Sample	72.8	56.8	0.03	173

Note. Response rates are defined as the number of household that completed the household questionnaire by web divided by the number of household that completed the household questionnaire in any other mode or that did not complete the household questionnaire (within each treatment group in the mixed-mode sample).

Table 16. Individual response rates between standard and targeted timing protocol, overall and by sample.

	Targeted	Standard	P	N
All	68.1	67.0	0.72	946
Original Sample	65.1	65.6	0.89	672
IP4 Refreshment Sample	75.7	70.1	0.30	274

Note. Response rates are defined as the number of full interviews by web divided by the number of full interviews in any other mode + the number of partial interviews + the number of non-respondents (within each treatment group in the mixed-mode sample).

References

Al Baghal, T. (ed.), Allum, N., Auspurg, K., Blake, M., Booker, C., Crossley, T.F., D'Ardenne, J., Fairbrother, M., Iacovou, M., Jäckle, A., Kaminska, O., Lynn, P., Nicoletti, C., Oldfield, Z., Pudney, S., Schnettler, S., Uhrig, S.C.N., & Winter, J. (contributors). (2014). "Understanding Society Innovation Panel Wave 6: Results from Methodological Experiments." in *Understanding Society Working Paper 2014-4*. University of Essex.

Al Baghal, T., Jäckle, A., Burton, J., & Lynn, P. (2016). Understanding Society The UK Household Longitudinal Study Innovation Panel, Waves 1-8, User Manual (pp. 1–144). Colchester Essex: Institute for Social and Economic Research University of Essex. Retrieved from https://www.understandingsociety.ac.uk/d/291/6849_ip_waves1-8_user_manual_June_2016.v3.1.pdf?1468270407

Atzmueller, C., & Steiner, P. M. (2010). Experimental Vignette Studies in Survey Research. *Methodology-European Journal of Research Methods for the Behavioral and Social Sciences, 6*(3), 128-138. doi: 10.1027/1614-2241/a000014

Bloom, D. E., Chatterji, S., Kowal, P., Lloyd-Sherlock, P., McKee, M., Rechel, B., *et al*. (2014). Macroeconomic implications of population ageing and selected policy responses. *The Lancet*. doi: 10.1016/s0140-6736(14)61464-1

Bowling, A. (2007). Aspirations for older age in the 21st century: what is successful aging? *Int J Aging Hum Dev, 64*(3), 263-297.

Bowling, A., & Dieppe, P. (2005). What is successful ageing and who should define it? *BMJ, 331*(7531), 1548-1551. doi: 10.1136/bmj.331.7531.1548

Brzozowski, M., & Crossley, T. F. (2011). "Measuring the Well-being of the Poor with Income or Consumption: A Canadian Perspective." *Canadian Journal of Economics* 44(1):88-106.

Burton, J. (2012). Mode effects in a self-completion instrument in Jonathan Burton (ed.) *Understanding Society* Innovation Panel Wave 4: Results from Methodological Experiments, *Understanding Society Working Paper Series* No. 2012-06, pp.35-40

CEC. (2009). *Dealing with the impact of an ageing population in the EU (2009 Ageing Report)*. Brussels: Commission of the European Communities.

Cernat, A., & Oberski, D. L. (forthcoming). Extending the within-persons experimental design: The multitrait-multierror (MTME) approach. In P. J. Lavrakas (Ed.), *Experimental Methods in Survey Research*. New York: John Wiley & Sons.

Christensen, K., Doblhammer, G., Rau, R., & Vaupel, J. W. (2009). Ageing populations: the challenges ahead. *The Lancet, 374*(9696), 1196-1208. doi: http://dx.doi.org/10.1016/S0140-6736(09)61460-4

Cosco, T. D., Prina, A. M., Perales, J., Stephan, B. C. M., & Brayne, C. (2013). Operational definitions of successful aging: a systematic review. *International Psychogeriatrics, FirstView*, 1-9. doi: doi:10.1017/S1041610213002287

Depp, C. A., & Jeste, D. V. (2006). Definitions and predictors of successful aging: A comprehensive review of larger quantitative studies. *American Journal of Geriatric Psychiatry, 14*(1), 6-20. doi: 10.1097/01.JGP.0000192501.03069.bc

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsely, T.L., Visscher, W., & Ezzati, T. M.. (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

Durrant, G. B., D'Arrigo, J., & Steele, F. (2011). Using paradata to predict best times of contact, conditioning on household and interviewer influence, *J.R. Statist. Soc. A*, 174, 1029-1049.

Fricker, S., Kopp, B. & To, N. (2015). "Exploring a Balance Edit Approach in the Consumer Expenditure Quarterly Interview Survey" in C. Carroll, T. Crossley and J. Sabelhaus (eds), Improving the Measurement of Consumer Expenditures, Chicago: University of Chicago Press, DOI:10.7208/chicago/9780226194714.003.0013.

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

Katz, S., & Calasanti, T. (2014). Critical Perspectives on Successful Aging: Does It "Appeal More Than It Illuminates"? *The Gerontologist*. doi: 10.1093/geront/gnu027

Kivimaki, M., & Ferrie, J. E. (2011). Epidemiology of healthy ageing and the idea of more refined outcome measures. *Int J Epidemiol, 40*(4), 845-847. doi: dyr114 [pii] 10.1093/ije/dyr114

Lynn, P. (2009). Sample design for Understanding Society, *Understanding Society Working Paper 2009-01*. Colchester: University of Essex.

Martin, P., Kelly, N., Kahana, B., Kahana, E., J. Willcox, B., Willcox, D. C., *et al*. (2014). Defining Successful Aging: A Tangible or Elusive Concept? *The Gerontologist*. doi: 10.1093/geront/gnu044

McLaughlin, S. J., Jette, A. M., & Connell, C. M. (2012). An Examination of Healthy Aging Across a Conceptual Continuum: Prevalence Estimates, Demographic Patterns, and Validity. *Journals of Gerontology Series a-Biological Sciences and Medical Sciences, 67*(7), 783-789. doi: 10.1093/gerona/glr234

Montross, L. P., Depp, C., Daly, J., Reichstadt, J., Golshan, S., Moore, D., *et al*. (2006). Correlates of self-rated successful aging among community-dwelling older adults. *American Journal of Geriatric Psychiatry, 14*(1), 43-51. doi: 10.1097/01.jgp.0000192489.43179.31

Rowe, J. W., & Kahn, R. L. (1997). Successful aging. [Article]. *The Gerontologist, 37*(4), 433-440.

Samphantharak, K., & Townsend, R. M. (2010). Households as Corporate Firms. New York: Cambridge University Press.

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

Strawbridge, W. J., Wallhagen, M. I., & Cohen, R. D. (2002). Successful Aging and Well Being: Self-Rated Compared With Rowe and Kahn. *The Gerontologist, 42*(6), 727-733. doi: 10.1093/geront/42.6.727

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

Tourangeau, R. & Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin* 133 (5): 859–883

UN. (2002). *Report of the Second World Assembly on Ageing A/CONF.197/9*. New York: United Nations.

von Faber, M., Bootsma-van der Wiel, A., van Exel, E., Gussekloo, J., Lagaay, A. M., van Dongen, E., *et al*. (2001). Successful aging in the oldest old: Who can be characterized as successfully aged? *Arch Intern Med, 161*(22), 2694-2700. doi: ioi00981 [pii]

Wagner, J. (2013), Adaptive contact strategies in telephone and face-to-face surveys, *Surv. Res. Meth.*, 7, 45-55.

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research, 38*(3), 505-520. doi: http://dx.doi.org/10.1016/j.ssresearch.2009.03.004

Young, Y., Frick, K. D., & Phelan, E. A. (2009). Can Successful Aging and Chronic Illness Coexist in the Same Individual? A Multidimensional Concept of Successful Aging. *Journal of the American Medical Directors Association, 10*(2), 87-92. doi: 10.1016/j.jamda.2008.11.003