Understanding Society

Working Paper Series

No. 2014-04

July 2014

Understanding Society Innovation Panel Wave 6:

Results from Methodological Experiments

Tarek Al Baghal (ed.)

Contributors: Nick Allum[1], Katrin Auspurg[2], Margaret Blake[3], Cara Booker[4], Thomas Crossley[1,8], Joanna D'Ardenne[3], Malcolm Fairbrother[5], Maria Iacovou[6], Annette Jäckle[4], Olena Kaminska[4], Peter Lynn[4], Cheti Nicoletti[7], Zoe Oldfield[8], Stephen Pudney[4], Sebastian Schnettler[2], S.C. Noah Uhrig[4], Joachim Winter[9]

[1] University of Essex, [2] University of Konstanz, [3] NatCen Social Research, London, [4] Institute of Social and Economic Research, University of Essex, [5] University of Bristol, [6] University of Cambridge, [7]University of York, [8] Institute of Fiscal Studies, [9] University of Munich

Understanding Society Innovation Panel Wave 6: Results from Methodological Experiments

Tarek Al Baghal (ed.)

Contributors: Nick Allum[1], Katrin Auspurg[2], Margaret Blake[3], Cara Booker[4], Thomas Crossley[1,8], Joanna D'Ardenne[3], Malcolm Fairbrother[5], Maria Iacovou[6], Annette Jäckle[4], Olena Kaminska[4], Peter Lynn[4], Cheti Nicoletti[7], Zoe Oldfield[8], Stephen Pudney[4], Sebastian Schnettler[2], S.C. Noah Uhrig[4], Joachim Winter[9]

[1] University of Essex, [2] University of Konstanz, [3] NatCen Social Research, London, [4] Institute of Social and Economic Research, University of Essex, [5] University of Bristol, [6] University of Cambridge, [7]University of York, [8] Institute of Fiscal Studies, [9] University of Munich

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 6 of the Innovation Panel in the spring of 2013.

IP6 was the second wave employing a mixed-mode design including an internet survey, and the third wave of the Innovation Panel to employ a mixed-mode design generally. IP2 had experimented with telephone interviewing in addition to face-to-face personal interviewing. Like IP5, IP6 uses a design in which a random two-thirds of households are allocated to a sequential mixed-mode design. The adults in these households were first approached by letter and email where possible and asked to complete their interview on-line. Those who did not respond on-line were then followed up by face-to-face interviewers. The remaining third of households were issued directly to face-to-face interviewers.

The methodological tests included an experiment testing the effects of changing the amount of incentives offered to respondents in advance of fieldwork on response rates, the use of targeted advance letters, the impact of changing the way a person responds from a personal to web interview impacts data quality, and the effect of answering questions using a computer or paper format. Further experiments examine the measurement of household energy use, the use of vignettes, the measurement of finger length as an indicator of health outcomes, the measurement of expenditures and consumption, the reliability of measures of change for disability status, the impact of being a panel member for a longer period on response choices, and the value of repeating questions about what format a respondent would like to respond to in future surveys.

Understanding Society Innovation Panel Wave 6: Results from Methodological Experiments

Tarek Al Baghal (ed.)

Contributors: Nick Allum[1], Katrin Auspurg[2], Margaret Blake[3], Cara Booker[4], Thomas Crossley[1,8], Joanna D'Ardenne[3], Malcolm Fairbrother[5], Maria Iacovou[6], Annette Jäckle[4], Olena Kaminska[1], Peter Lynn[4], Cheti Nicoletti[7], Zoe Oldfield[8], Stephen Pudney[4], Sebastian Schnettler[2], S.C. Noah Uhrig[4], Joachim Winter[9]

[1] University of Essex, [2] University of Konstanz, [3] NatCen Social Research, London, [4] Institute of Social and Economic Research, University of Essex, [5] University of Bristol, [6] University of Cambridge, [7]University of York, [8] Institute of Fiscal Studies, [9] University of Munich

Abstract

This paper presents some preliminary findings from Wave 6 of the Innovation Panel (IP6) of *Understanding Society*: The UK Household Longitudinal Study. *Understanding Society* is a major panel survey in the UK. In March 2013, the sixth wave of the Innovation Panel went into the field. IP6 used a mixed-mode design, using on-line interviews and face-to-face interviews. This paper describes the design of IP6, the experiments carried and the preliminary findings from early analysis of the data.

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

JEL classification: C80, C81, C83

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

Table of Contents

1. Introduction

This paper presents early findings from the sixth wave of the Innovation Panel (IP6) of *Understanding Society*: The UK Household Longitudinal Study (UKHLS). *Understanding Society* is a major panel survey for the UK. The first four waves of data collection on the main sample have been completed, and fifth and sixth waves are currently in the field. The data from the first three waves of the main samples are available from the UK Data Archive, and the fourth will be available towards the end of 2014. Data from a nurse visit to collect bio-markers from the general population sample and the British Household Panel Survey (BHPS) are also available. Data for the first six waves of the Innovation Panel are available from the UK Data Service[1].

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

The second wave of the Innovation Panel (IP2) was carried out in April-June 2009, the third wave (IP3) in April-June 2010 and the fourth wave in March-July 2011. The fourth wave of the Innovation Panel (IP4) included a refreshment sample of 465 responding households. In March 2012, IP5 was fielded, with part of the samples conducting the survey via the internet, while others continued in an interviewer-administered survey. Working Papers which cover the experimentation carried out in all six innovation panels are available from the

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

Understanding Society website.[2] The data from the first six waves of the innovation panel are held at the UK Data Service. This paper describes the design of IP6, the experiments carried and some preliminary findings from early analysis of the data. Section 2 outlines the main design features of *Understanding Society*. Section 3 describes the design and conduct of IP6. Section 4 then reports on the experiments carried at IP6.

2. *Understanding Society*: the UKHLS

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

The design of the main-stage of *Understanding Society* is similar to that of the British Household Panel Survey (BHPS) and other national panels around the world. In the first 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

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

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

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

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

3. Innovation Panel Wave 6: Design

IP6 employed a mixed-mode design, as in IP5. At IP5 and IP6 the modes which were mixed were on-line (CAWI) and face-to-face (CAPI) interviewing. In IP5, a random selection of two-thirds of households was allocated to the mixed-mode design ("WEB") with the remaining third of households allocated directly to face-to-face interviewers ("F2F"). This sample allocation was maintained at Wave 6. Additionally in Wave 6, if individuals had not participated by the end of the fieldwork period, they were assessed for inclusion in a final telephone interviewing (CATI) phase. The CAWI option was also available during this phase.

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

At the end of two weeks, all adults who had not completed their interview were allocated to face-to-face interviewers, but could still enter the web survey instead if they desired. Adults who had started their interview on-line, but not reached the 'partial interview' marker, were issued to face-to-face interviewers. The interviewers were able to re-start the interview at the place at which the respondent had stopped. Also at this point the remaining third of households, those in the F2F group, were issued to interviewers. The two-week WEB-only period before face-to-face fieldwork was implemented so that the face-to-face interviewers would have their full allocation at the start of their fieldwork, rather than having non-responding WEB individuals being passed to them during the fieldwork period. This was done to allow the face-to-face interviewers to work more efficiently.

The WEB-only period ran from 22nd March to 7th April. The face-to-face fieldwork started 8th

April and ran until 1st July. During this period the CAWI survey remained 'open' so that WEB individuals could complete their interview on-line during this fieldwork period. The mop-up follow-up phase with those not responding in both the WEB and F2F versions, conducted through CATI with CAWI available was from 4th July to 29th July.

Prior to the survey going into the field there were eleven one-day briefings for the interviewers. The briefings were conducted by NatCen researchers, with staff from ISER contributing to provide information about the study and to talk in more detail about the experiments. The locations of the briefings gave a wide geographic spread: London (six briefings), Leeds, Bristol, Derby, Manchester and Edinburgh. In total, 121 interviewers were briefed to work on IP6. A debrief also took place in July with a selection of interviewers from different areas. All interviewers working on the survey were provided with feedback forms and were asked to fill and return them to the NatCen operations office at the end of fieldwork. The questionnaires used at IP6 are available from the *Understanding Society* website.[4]

a. Call for experiments

IP6 was the fourth time the Innovation Panel was open for researchers outside the scientific team of *Understanding Society* to propose experiments. A public call for proposals was made on 21st March 2012 with a deadline of 15th May. Fourteen proposals were received with eight being accepted, plus four carried over from IP5, for a total of twelve being included in IP6. The fourteen new submissions came from within ISER (seven), ISER in collaboration with other researchers (five) and from outside ISER completely (two). Of those that were external to ISER, one was from the United States and the others were all from institutions in the UK or were collaboration between UK and international institutions. The fourteen proposals were reviewed by a panel which included two ISER-based members of the *Understanding Society* scientific leadership team, and two members of the Methodology Advisory Committee to *Understanding Society* who were external to ISER. In addition to those experiments which were accepted through the public call, there were a number of core experiments which the Understanding Society senior leadership team wanted to run. These

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

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

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

b. Sample

The sample issued for IP6 included the original sample and the refreshment sample which had first been interviewed at IP4. The original sample at IP6 comprised those households who had responded at IP5, plus some households which had not responded at IP5. 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 993 original and 461 refreshment sample households issued at IP6. Of these, 135 original and 47 refreshment sample households had not responded in IP5.

As discussed above, around two-thirds of the sample were allocated to the mixed-mode design in IP6, in which sample members would be approached by letter and email (where possible) to complete their interview on-line. This experimental allocation covered both the original and refreshment sample. The table below shows the allocation to mode design by sample type for those included in the issued samples in IP6.

Table 1: Allocation to mode design by sample type

	Original Sample	Refreshment Sample	Total
CAPI only	342 34.4%	167 36.2%	509 35.0%
Mixed-mode (CAWI+CAPI)	651 65.6%	294 64.8%	945 65.0%
Total	993	461	1,454

c. Questionnaire design

The questionnaire at IP6 followed the standard format used in the previous Innovation Panels as well as the main-stage of *Understanding Society*. The interview included:

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

Unlike some previous IPs, IP6 did not include audio recording of any portions of the interview.

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

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

about who was responsible for paying household bills was included; the person or people indicated as responsible were routed first to the household questionnaire and then to the individual questionnaire.

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

d. Response rates

This section sets out the response rates for IP6 as a whole. Section 4b describes the effect of incentives on response rates. Table 2 sets out the response rates for eligible households for the refreshment sample and the original sample. Immediately following, Table 3 separates out the response rate for households that had responded at IP5 and those that had not. In all tables, cells present both the percentage and the number of cases this percentage represents, while the bottom row presents total number of cases.

Table 2. Household response at IP6

	Original Sample	Refreshment Sample	Total
Responding	81.7% 811	82.7% 381	81.2% 1192
Non-contact	1.9% 19	2.6% 12	2.1% 31
Refusals	8.7% 86	10.2% 47	9.2% 113
Other non-responding	7.8% 77	4.6% 21	6.7% 98
Total	993	461	1454

Table 3. Household response at IP6 by IP5 outcome

	Original Sample		Refreshment sample	
	IP5 Responding	IP5 Non-Responding	IP5 Responding	IP5 Non-Responding
Responding	87.5% 751	44.4% 60	88.2% 365	34.0% 16
Non-contact	1.2% 10	6.7% 9	2.2% 9	6.4% 3
Refusals	6.4% 55	23.0% 31	7.5% 31	34.0% 16
Other non-responding	4.9% 42	25.9% 35	2.2% 9	25.5% 12
Total	858	135	414	47

There is not a significant difference identified in response outcomes overall by original or refreshment sample classification. The original sample response rate is also somewhat higher for IP6 (81.7%) than for IP5 (75.5%). Households who had responded at IP5 were, not surprisingly, more likely to respond at IP6. Original sample households that did not respond in IP5 were somewhat more likely to respond in IP6 than refreshment sample households that did not respond in IP5. Similarly, non-responding original sample households were less likely to refuse at IP6 than the corresponding households from the refreshment sample (noting the small numbers of households). Otherwise, household outcomes were similar across samples regardless of previous wave outcome.

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

Table 4. Household response at IP6 by CAPI or Mixed-Mode Design

	Total		Original Sample		Refreshment sample	
	F2F	MM	F2F	MM	F2F	MM
Responding	82.1% 418	81.9% 774	82.2% 281	81.4% 530	82.0% 137	83.0% 244
Complete HH	61.5% 313	65.0% 614	61.7% 211	63.3% 412	61.1% 102	68.7% 202
Partial HH	20.6% 105	16.9% 160	20.5% 70	18.1% 118	21.0% 35	14.3% 42
Non-contact	2.2% 11	2.1% 20	2.1% 7	1.8% 12	2.4% 4	2.7% 8
Refusals	9.0% 46	9.2% 87	8.5% 29	8.8% 57	10.2% 17	10.2% 30
Other non-responding	6.7% 34	6.8% 64	7.3% 25	8.0% 52	5.4% 9	4.1% 12
Total	509	945	342	651	167	294

There is little difference between the CAPI-only and mixed-mode designs in overall response rate (combining complete and partial response). The only apparent difference is that for complete household response, with the mixed-mode design having somewhat higher percentage (65.0%) than the CAPI-only design (61.5%). This finding is opposite of that in IP5, where response rates for the mixed-mode overall were lower. This difference in complete household response is due largely to the significantly higher percentage for the mixed-mode design (68.7%) than the CAPI-only design (61.1%) in the refreshment sample.

Turning from the household to the individual, Table 5 presents individual re-interview rates. There were 2,023 individual respondents aged 16 or older fully interviewed in IP6. As with household response, there is not a significant effect identified for the different samples.

However, the refreshment sample has a somewhat higher percentage of personal interviews (73.9%) than the original sample (71.4%), whereas there are a slightly higher percentage of proxy interviews in the original sample (4.8% v. 3.9%).

Table 5. Individual re-interview response at IP6

	Original Sample	Refreshment Sample	Total
Personal Interview	71.4%	73.9%	72.2%
	1,356	667	2,023
Proxy Interview	4.8%	3.9%	4.5%
	91	35	126
Non-contact	5.7%	5.5%	5.7%
	109	50	159
Refusal	13.6%	13.0%	13.4%
	259	117	376
Other non-response	4.5%	3.8%	4.3%
	85	34	119
Total	1,900	903	2,803

The individual-level response rates for continuing and refreshment samples in IP6 across survey mode designs are shown in Table 6 below. There were few partial interviews (1.4% overall) and almost all occurred in the web survey while only 3 occurred in the CAPI-only mode. Noting this, these outcomes are included as personal interviews. Overall, the mixed-mode design has somewhat higher individual re-interview rates, again, contrary to the findings in IP5. The percentage for personal interviews is also somewhat higher for both designs in the refreshment sample. Refusals are higher in the mixed-mode design relative to the CAPI-only; this difference is most marked in the refreshment sample. The percentages for proxy interviews are consistently higher in the CAPI-only design, and in the CAPI-only design, proxy interviews are relatively greater in the original sample.

Table 6. Individual re-interview response at IP6 by mode

	Total		Original Sample		Refreshment sample	
	F2F	MM	F2F	MM	F2F	MM
Personal Interview	70.7%	73.0%	70.0%	72.1%	71.9%	75.0%
	686	1,337	446	910	240	427
Proxy Interview	7.2%	3.1%	8.0%	3.2%	5.7%	2.8%
	70	56	51	40	19	16
Non-contact	5.7%	5.7%	5.2%	6.0%	6.6%	4.9%
	55	104	33	76	22	28
Refusal	12.1%	14.1%	12.7%	14.1%	10.8%	14.2%
	117	259	81	178	36	81
Other non-response	4.4%	4.2%	4.1%	4.7%	5.1%	3.0%
	43	76	26	59	17	17
Total	971	1,832	637	1,263	334	569

Since IP6 introduced a "mop-up" phase where respondents were contacted by telephone to complete the survey, as well as opening the web version to anyone not yet responding, Table 7 presents the mode actually responded to for all respondents.[6] Not surprisingly, almost all of the CAPI-only assigned respondents completed survey in a face-to-face setting. While the majority of respondents assigned to the mixed-mode design completed the web version, a sizable minority responded when an interviewer approached them at home. However, among those assigned to the mixed-mode design, significantly more respondents in the refreshment sample responded to the web version than those in the original sample.

[6] Six respondents did not have a final mode recorded in the data. Three were originally assigned to the CAPI-only and three to the mixed-mode design.

Table 7. Survey mode of response

	Total		Original Sample		Refreshment sample	
	F2F	MM	F2F	MM	F2F	MM
Face-to-Face	97.1%	38.5%	97.8%	42.8%	95.8%	29.3%
	663	514	436	389	227	125
Web	2.2%	60.5%	1.4%	56.8%	3.8%	68.3%
	15	807	6	516	9	291
Telephone	0.7%	1.0%	0.9%	0.3%	0.4%	2.4%
	5	13	4	3	1	10
Total	683	1334	446	908	237	426

A small number of CAPI-only respondents ended up responding to the web during the mop-up period, more than to the telephone invitation. Slightly more respondents in the mixed-mode design responded to the telephone interview than those assigned to the CAPI-only design, but the numbers too small to make conclusions. However, taken together, it is clear that the mop-up phase added a number of respondents who otherwise would have been treated as non-productive outcomes.

4. Experimentation in IP6

There were a number of experiments carried on IP6 covering both fieldwork procedures and measurement in the questionnaire. There were some new experiments and some which were the longitudinal continuation of experiments carried at previous waves of the IP. This section outlines the experiments carried at IP6; 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.

a. Assessing the Feasibility of More Precisely Measuring Household Energy Consumption (Malcolm Fairbrother)

This experiment aimed to assess the feasibility of measuring energy use within the households. Households were randomly assigned to four treatment conditions, based on two crossed binary treatments: (i) The advance letter for adults in a randomly-selected half of households included a paragraph which mentioned the proposal to collect meter-readings. The advance letter for adults in the other half of households did not mention the meter-reading collection. (ii) Half of the households were asked only for an odometer reading from the household's most used vehicle, while half were asked for that, plus readings from their gas and/or electric meter(s). Given differences in self-selection with respect to survey mode, the random assignment to these four treatment conditions did not achieve balance across survey modes; the number of face-to-face households by treatment group, for example, ranged from 179 to 195.

Of the 1189 households, 824 (69%) provided at least one valid gas, electricity and/or odometer reading (405 of these were only an odometer reading). Approximately one month after the end of fieldwork for IP6, a postal questionnaire was sent to households who had given a meter-reading. The questionnaire asked the household to give another meter-reading, enabling researchers to calculate the energy use between the date of interview and the date of the second meter-reading. A reminder letter was sent to those households who had not returned their questionnaire after two weeks. Two weeks after that, non-responding households were contacted by telephone and the meter-readings collected, if possible, by telephone. The follow-up survey was only requested of those who gave a valid response to one or more of these questions, and these households were only asked a repeat of what they had answered initially. Of the 824, 672 (82%, or 57% of 1189) completed the follow-up survey.

Of the 1189 households, setting aside a very small number of missing values, 1173 reported having electricity, 1030 gas, 55 oil, and 99 some other kind of fuel in their homes; 959 households reported having one or more vehicles. At the main IP6 interview, neither the treatment of receiving an advance letter, nor the treatment of being asked only for an odometer reading rather than odometer plus gas/electricity meter, appears to have made any

difference to compliance with the request to provide an odometer reading. 76% of households with at least one vehicle complied with this request (or 80% if disregarding DKs and some households coded as Inapplicable).

In terms of data quality, those asked only for an odometer reading (and not also gas/electricity readings) were *less* likely to provide a precise figure (44% compared to 50%), and more likely to report an estimate. The need to collect one outside piece of information (a meter reading) may have led some respondents to make the extra effort also to record their precise odometer reading. The advance warning letter made no difference.

Survey mode made some difference; web respondents were slightly less likely to provide an odometer reading than face-to-face respondents (73% compared to 86%), and were also slightly less likely to provide a precise reading rather than an estimate (43% rather than 49%). (Too few households responded by telephone for any comparison to be meaningful.)

With respect to gas, depending on how missing data are treated, 70% to 80% of households proved willing to provide a gas meter reading, and being warned ahead of time made little to no difference. Almost all provided a precise number rather than an estimate (presumably because few people have any idea what their gas meter reads unless they look). Receiving an advance warning letter made no difference to the probability of compliance. Respondents differed substantially in terms of non-response depending on survey mode, with face-to-face respondents by far the likeliest to provide a gas reading, compared to telephone and web respondents (84% rather than 61% and 67%, respectively, with the number of telephone respondents being very small). Survey mode made little difference to the probability of providing an estimated figure, however.

Much the same held for electricity, though for electricity only 60% to 70% of households proved willing to provide an electricity meter reading. Again being warned ahead of time made little to no difference, and almost all provided a precise number rather than an estimate (also irrespective of being warned ahead of time). As for gas, web respondents were much less likely to provide an electricity reading compared to face-to-face respondents (64% rather

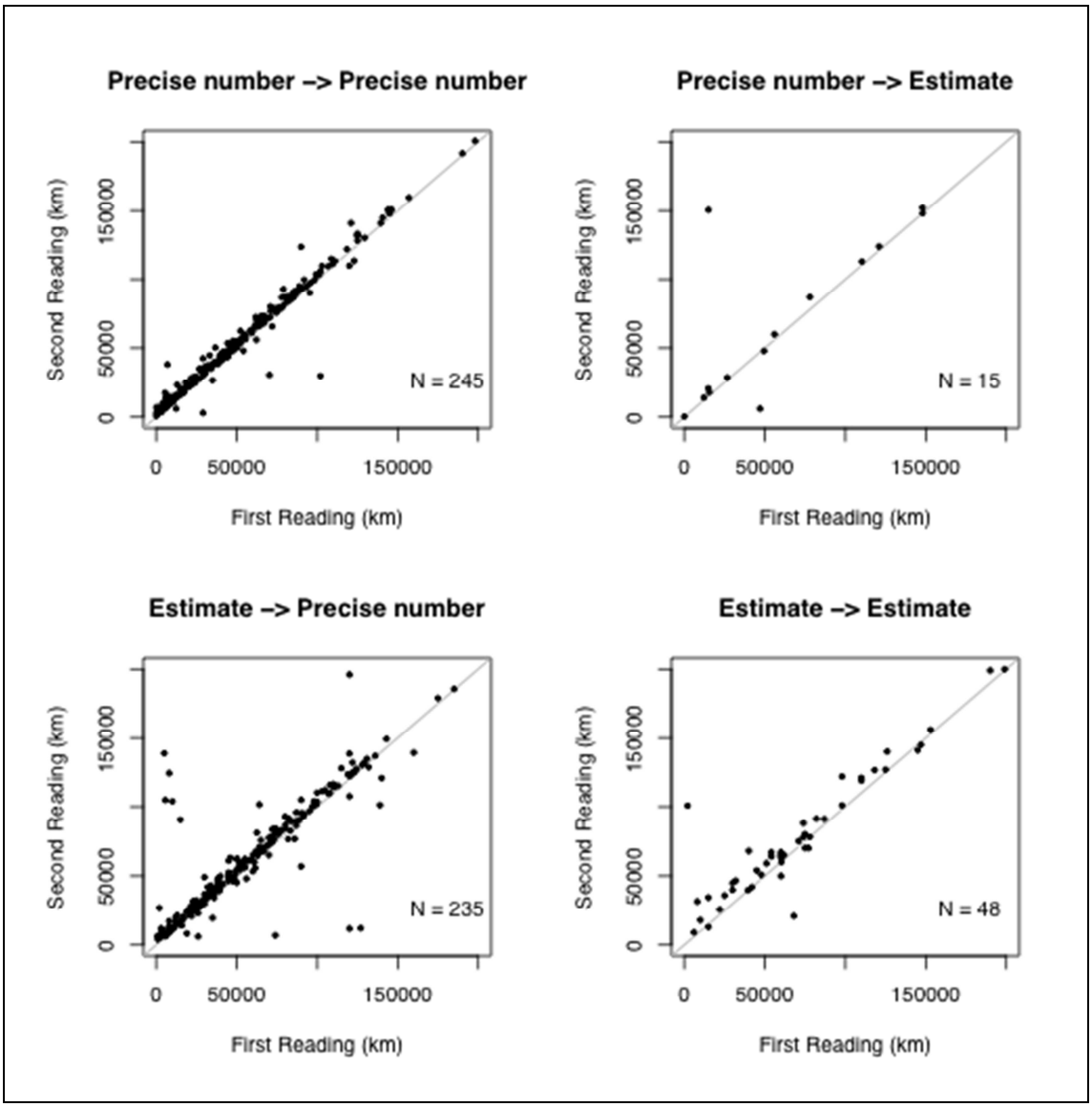
than 90%), with survey mode again making little difference to the (inevitably very low) probability of providing an estimated rather than precise figure.

Overall, then, a warning in the advance letter does not appear to make much difference in the collection of these data. On the other hand, survey mode does make a difference: for collecting information about gas and electricity use, item non-response is high for web respondents, and low for face-to-face respondents.

Turning to the second-stage data collection, respondents were asked whether their address and (if appropriate) vehicle were the same as those recorded in the first stage. Despite this check, however, in a non-trivial number of cases, odometer and/or meter readings were lower at the time of the second reading compared to the first. This occurred for 11% of households who reported odometer readings at both stages and passed a check that the vehicle was unchanged; the same occurred for 7% of households with respect to electricity and gas meter readings (and where their address should have been unchanged). In other cases, the figures were dramatically, unrealistically higher. In both types of instances, the quality of the data would clearly appear to be suspect. An error must have been made at one or both stages, or the vehicle must have changed—yet such a change was not reflected in the data.

Those cases aside, however, the majority of respondents who gave readings at the first stage also complied at the second stage. Interestingly, for odometer readings, almost all second-stage readings were precise figures, not estimates—in contrast (as discussed above) to the first stage (see Figure 1 below). This would suggest that the follow-up cards encouraged respondents to go and look at their odometers, whereas in the initial web or face-to-face interviews they tended just to estimate. Large majorities of the respondents used the mail-back cards to report their odometer and meter readings, though there was also a significant minority who responded only when prompted by telephone.

Figure 1: Odometer readings, by type of reading at each stage (estimated or precise).



b. Change in Respondent Incentives (Peter Lynn)

At IP6, as at previous waves, sample members were sent an unconditional incentive with the advance letter notifying them of the upcoming wave of data collection. The value of the incentive, which was in the form of a voucher redeemable for cash at any Post Office, was either £10 or £30. Additionally, some of those sent £10 in the mixed mode treatment group were also promised an additional £20 for each adult household member conditional on all adult household members taking part online within two weeks of receiving the survey

invitation. For some sample members, this was the same level of incentive that they had received at IP5, but for most this represented a change. In some cases, the incentive level was increased while in others it was reduced.

In the CAPI-only part of the sample, all sample members were provided a £10 incentive. Amongst original sample members this represented either an increase from £5 at IP5 or the same that they had received at IP4. Amongst IP4 refreshment sample members, this represented a reduction from £20 or £30 at IP4 or the same that they had received at IP4. Amongst original sample members, IP6 response rate was slightly higher for those for whom the incentive represented an increase, though the difference did not reach statistical significance (Table 8). Amongst refreshment sample members, the opposite was found: response rate was higher for those for whom the incentive represented a reduction, though again the difference did not reach statistical significance. Differences between treatment groups are even smaller if analysis is restricted to previous wave respondents.

Table 8: Response rates by change in incentive level, sample origin, and previous wave response status; CAPI-only sample

	Original sample		IP4 refreshment sample		
Incentive level at IP5 and IP6	(a) £5 → £10	(b) £10 → £10	(c) £10 → £10	(d) £20 → £10	(e) £30 → £10
All issued to field	79.2% (n=337)	73.9% (n=299)	72.1% (n=86)	75.6% (n=127)	80.2% (n=121)
IP5 respondents	88.2% (n=245)	85.2% (n=209)	87.7% (n=57)	89.9% (n=89)	87.4% (n=95)

All issued to field: (a) v (b) P=0.11; (c) v (d) P=0.56; (c) v (e) P=0.18; (d) v (e) P=0.39
IP5 respondents: (a) v (b) P=0.35; (c) v (d) P=0.68; (c) v (e) P=0.95; (d) v (e) P=0.59

In the mixed mode part of the sample, there were three different incentive treatments at IP6, as described above. Original sample members could have received either £5 or £10 at IP5 and thus, there are six treatment combinations across the two waves, of which five represent increases of different amounts and one represents a constant treatment (of £10). Amongst those who received £5 at IP5, either of the higher levels incentives resulted in a significantly higher response rate at IP6 than the £10 incentive (Table 9). Amongst those who received £10

at IP5, only the £30 incentive at IP6 resulted in a significantly higher response rate than the £10 incentive. Amongst those who received £10 at IP6, the IP6 response rate did not differ between those who received £5 and those who received £10 at IP5. The same was true for sample members receiving each of the other two levels of incentives at IP6. Broadly speaking, these results can be characterised as indicating that:

- conditional on the level of incentive at IP5, higher levels of incentives at IP6 resulted in higher response rates;
- conditional on the level of incentive at IP6, the level of incentive at IP5 did not affect the response rate at IP6.

These findings are consistent with a hypothesis that the effect on response propensity is driven by the current level of the incentive, not by the change in level from one wave to the next.

Table 9: Response rates by change in incentive level, and previous wave response status; original sample, web sample

Incentive level at IP5 and IP6	(a) £5 → £10	(b) £5 → £10+£20	(c) £5 → £30	(d) £10 → £10	(e) £10 → £10+£20	(f) £10 → £30
All issued to field	64.8% (n=230)	75.5% (n=224)	78.1% (n=233)	67.8% (n=177)	75.5% (n=196)	81.8% (n=203)
IP5 respondents	83.9% (n=137)	87.9% (n=149)	93.3% (n=150)	80.3% (n=127)	84.4% (n=141)	88.9% (n=135)

All issued to field: (a) v (b) P=0.01; (a) v (c) P=0.001; (b) v (c) P=0.50; (d) v (e) P=0.10; (d) v (f) P=0.002; (e) v (f) P=0.13; (a) v (d) P=0.52; (b) v (e) P=0.99; (c) v (f) P=0.34
IP5 respondents: (a) v (b) P=0.33; (a) v (c) P=0.01; (b) v (c) P=0.11; (d) v (e) P=0.38; (d) v (f) P=0.05; (e) v (f) P=0.27; (a) v (d) P=0.44; (b) v (e) P=0.39; (c) v (f) P=0.19

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

This experiment used reactive dependent interviewing to investigate the measurement of change in self-assessed measures of long-standing illness or disability. Waves 1-4 of the IP contain the same question asking whether the respondent is troubled by a long-standing (at least 12 months) illness, disability or infirmity. There has been some experimental variation

in wording but, comparing individual responses from the same question design in successive waves, we find high rates of transition: exit rates from disability of 29% (IP1→IP2 and IP2→IP3) and 33% (IP3→IP4); and entry rates of 18% (IP1→IP2), 24% (IP2→IP3) and 14% (IP3→IP4). These seem implausibly high for the general population, given the "long-standing" qualifier.

In substantive research, questions of this kind are often used to construct variables identifying people suffering ill-health, and to identify adverse health events. They are important in epidemiology, but are also widely used as explanatory variables in survey-based research in many other fields, including labour economics, income distribution, wellbeing and poverty analysis, tax-benefit modelling and planning of public services. If there proves to be a great deal of spurious "churning" in responses, this will have serious implications for a great deal of important empirical research.

The question is important in its own right, but it is also used in *Understanding Society* and the *Family Resources Survey* (FRS) as a filter that precedes a question inviting respondents to report specific difficulties with a set of 11 specific activities of daily life (ADLs). Responses to this second stage question are often used to construct empirical measures of the severity of disability, based on the number and types of difficulties that a person reports. Measures of this kind have been influential in academic and policy-related research on disability. Examples include the Wanless Review (2006) and the 2009 Green Paper on social care (Department of Health 2009), both of which reached conclusions about the targeting of support for disabled people, based on these measures.

If the initial filter question is unreliable, as seems possible given the high rate of churning, then measures of disability constructed from the reported difficulties with ADLs may be systematically biased, even if errors in responses to the filter question are purely random. This is because of the asymmetric question structure: a random "false negative" response bars entry to the ADL question and thus prevents reports of any difficulty, but a false positive does not necessarily lead to an offsetting over-estimate of ADL difficulties. This bias may have serious implications for evidence-based design of disability policy, since it could lead to

underestimation of the prevalence of disability and the accuracy of targeting of public support for disabled people.

Objectives

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

Experimental design

Sample members were randomly (by household) allocated to one of three experimental groups.

Group A (quarter of the sample):

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

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

If HEALTH=yes:

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

1 Mobility (moving around at home and walking)

2 Lifting, carrying or moving objects

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

4 Continence (bladder and bowel control)

5 Hearing (apart from using a standard hearing aid)

6 Sight (apart from wearing standard glasses)

7 Communication or speech problems

8 Memory or ability to concentrate, learn or understand
9 Recognising when you are in physical danger
10 Your physical co-ordination (e.g. balance)
11 Difficulties with own personal care (e.g. getting dressed, taking a bath or shower)
12 Other health problem or disability
96 None of these

Group B (quarter of the sample):
Everyone was asked the ADL question; the HEALTH filter question was not asked.

Group C (half the sample):
Everyone was asked the HEALTH question about long-standing health conditions. Respondents who gave a different answer from the previous wave were asked a follow-up question about the reasons for the change:
Can I just check, our records show that last time when we interviewed you on [ff_intdate] , {you had a / you did not have any} long-standing illness or disability. Is there an error in our records, or {do you no longer have this condition / is this a new condition}?

Everyone in this group was also asked the ADL question, but at a later point in the questionnaire. The experiment is being repeated in IP7.

Results
1,293 respondents answered the filter question at both wave 5 and wave 6. A third of these (426) reported an initial health condition or disability at wave 5, of whom 80 reported no condition at wave 6 – an exit rate of 19%. Among the 867 respondents at wave 5 who reported no long-standing condition, 123 reported such a condition at wave 6 – an entry rate of 14%. These entry and exit rates are lower than the corresponding rates in some earlier waves, but they remain implausibly high.

Table 10 documents the explanations respondents gave for changes in their long-term illness or disability status. Of the 45 respondents no longer reporting a long-term problem, 11

confirmed that they no longer had the condition; 29 said they still had the same condition, but that it was not as bad now, or medication/treatment was more effective, or it was less of a problem because their activities had changed. Only 5 respondents said there was an error in their data from the previous interview or other reason for the change in their health status.

Table 10: Reasons for changes in long-term health status

Reasons for no longer reporting long-term illness/disability	N
There is an error in the records	4
I still have the same health condition but it is not as bad now	8
I still have the same health condition but treatment or medication is effective now	15
The condition is much the same as last year, but my activities have changed, so it is less of a problem now	6
I no longer have this health condition	11
Other reason	1
Total	45
Reasons for reporting new long-term illness/disability	**N**
There is an error in the records	20
I had the same health condition but it is worse now	8
I had the same health condition but treatment or medication is less effective now	3
The condition is much the same as last year, but my activities have changed, so it is more of a problem now	3
This is a new health condition	29
Other reason	5
Total	68

Measuring the onset of new long-term health conditions seems more problematic. Of the 68 respondents who gave an explanation for reporting new long-term health problems 29 confirmed that they had a new health condition; 14 said they had had the condition previously but that it was worse now, or that the treatment or medication was less effective, or that it was more of a problem now because their activities had changed. However, 25 respondents said there was an error in the data from their previous interview, or another reason.

Disability rates based on the questions about Difficulties with Activities of Daily Life tended to be higher if everyone was asked this question, than when respondents were only routed into this question if they reported a long-term illness or disability. Among IP6 respondents the rates were 27.4% versus 23.3% (p=0.180), among respondents also interviewed in the previous wave the rates were 30.1% versus 24.2% (p=0.088). Correspondingly the mean numbers of activities that respondents had difficulty with tended to be somewhat higher if everyone was asked the question than when it was routed (0.59 versus 0.49 for IP6 respondents, p=0.230, and 0.64 versus 0.51 for respondents in both IP5 and IP6, p=0.133).

d. Panel Conditioning and Social Desirable Responding: Effects on self-reported height and weight (S.C. Noah Uhrig)

Validation work on self-reported height and weight in surveys suggests that both are consistently biased toward cultural ideals (for a review, see Rowland, 1990, see also Spencer et al., 2002). Weight is often under-reported, particularly amongst those who are heavier whilst height can also be over-reported amongst those who are short or under-reported amongst those who are tall. Such biases often lead to misclassifications of relative weight – i.e., underweight, normal weight, overweight or obese -- a much studied variable in epidemiology and other disciplines. Little is known, however, about the longitudinal measurement properties of self-reported factual data, like height and weight, particularly in the context of clear social desirability bias in self-reports. Adopting the notion that survey content informs and encourages panel respondents to trust the survey enterprise and hence report more accurately (Waterton and Lievesley 1989), this study investigates whether questionnaire content from a prior wave reduces the likelihood of observing these biases in height and weight self-reports.

***Experimental Design*.**

At Wave 1, half of IP respondents were exposed to height and weight questions while the other half of the sample was not. At Wave 2, the entire sample was asked their height and weight. This approach was replicated twice more with identical allocations for each pair of waves IP3 and IP4, and IP5 and IP6. Thus, content was varied across six waves to be either annual or biannual; the portion of the sample receiving height and weight content annually is

treated as "conditioned" while biannual content is "not conditioned". Households within PSUs were randomly allocated to one or the other treatment such that all respondents within a given household received the exact same experimental treatment.

Background and Hypotheses.

Validation of self-reported weight against anthropometric measurement finds that weight is systematically underreported (Dekkers et al., 2008, Spencer et al., 2002, Borkan et al., 1983). Underreporting is consistently greater among those who are heavier and by women (Rowland, 1990, Stewart et al., 1987, Palta et al., 1982). Both Spencer et al., (2002) and Rowland (1990) find that the extent of under-reporting of weight increases with increasing respondent weight, more so for women than for men. The margin of error for women is typically twice that for men at the heaviest weights within sex. Rowland (1990) also finds that underweight men over-report their weight. Validation of self-reported height typical finds that height is over-reported, though generally by small margins (Rowland, 1990, Spencer et al., 2002, Dekkers et al., 2008). As with weight, misreports seem to be associated with both gender and true value. Typically, greater over-reports are observed amongst shorter men (Rowland, 1990). Though generally of a small magnitude, Spencer et al., (2002) find that men's overestimates are nearly twice that of women.

Height and weight are not validated in the UKHLS IP, nevertheless panel conditioning effects on self-reported height and weight can still be investigated. If conditioning reduces socially desirable responding through the inculcation of trust, one would expect the response distribution for both height and weight to be affected for men and women in ways suggesting less socially desirable responding.

> *H1: Conditioned respondents should report heights and weights systematically opposed to the biases identified by validation work. Unconditioned respondents should confirm known biases, or rather show no conditioning effects.*

Wilfully providing inaccurate information is one socially desirable response strategy, but there are others (Tourangeau et al., 1997, Tourangeau et al., 2000). First, item non-response

is a common method to avoid providing information which is unflattering or otherwise highly sensitive (Moore et al., 1999, Kennickell, 1996). Panel conditioning research often finds that non-response decreases over waves of data collection (Traugott and Katosh, 1979, Bailar, 1989, Cantor, 1989, Porst and Zeifang, 1987, Sturgis et al., 2009, Waterton and Lievesley, 1989).

> *H2: Conditioned men and women should be less likely to item non-respond for both height and weight than unconditioned men and women.*

A third socially desirable response strategy could be to provide round numbers, i.e., "digit preference", particularly among heavier respondents. In his study of U.S. men and women, Rowland (1990) found that 60 percent expressed a digit preference -- i.e., a numeric value ending in a 0 or 5 -- when reporting weight in imperial units. Digit-preference was more common among women and heavier respondents and those expressing a digit preference were significantly less accurate than those who did not (Rowland 1990).

> *H3: Conditioning influences digit preference such that conditioned respondents should be less likely to provide rounded values for height and weight compared to unconditioned respondents.*

Variables.

Respondents were asked their height without shoes which could be reported either in metric or imperial units. Respondents were also asked their weight without clothes which could similarly be reported in either metric or imperial units. Few respondents reported in metric, therefore the analysis is limited to only those reporting both height and weight in imperial units. The weight question was followed-up with an indicator of whether the reported value is an estimate or not, and when the respondent most recently weighed themselves. All women currently pregnant were excluded from the analysis.

Rounding in weight reporting was indicated by whether the respondent provided an answer that was a full or half-stone (e.g., "12 stone" or "12-1/2 stone" rather than "12 stone 3 pounds", etc,...).

***Methods*.**

Validation results suggest that the biases in self-reports of weight and height differ over the range of the distribution; however, linear regression models the mean values of the response variables conditional on a given set of predictors. Quantile-regression models the conditional response distribution rather than the mean, i.e., a specified percentile or percentiles of a continuous response variable conditioned on a set of covariates (Koenker and Bassett, 1978). For this reason, quantile-regression is more appropriate for examining the effects of panel conditioning on the underlying distribution of responses to height and weight questions. In addition to quantile-regression, routine logit and multinomial logit models are appropriate to examine other effects. In all models, respondent age and education are controlled. Education was measured in terms of highest qualifications obtained, categorised as into four groups: University degree or higher, or an equivalent; Completion of compulsory schooling or its equivalent, including those staying on until age 18; all other qualifications not elsewhere classified, including foreign degrees; and no reported qualifications at all.

***Results*.**

Table 11, for men, and Table 12, for women, show estimates of panel conditioning effects on quartiles of self-reported weight. Results for men suggest no support for H1 such that conditioning has no statistically significant effect on reported weight. Results for women suggest some initial support for H1: conditioned women in the upper 75th percentile routinely report heavier weights by about 1/2-stone. This result, however, is not replicated at Wave 4 nor at Wave 6 though conditioned women in this upper quartile are likely to report heavier weights than unconditioned women.

Table 11. Quantile-regression of conditioning, rounding and recent weighing on men's self-reported weight.

	Wave 2			Wave 4			Wave 6†		
	I	II	III	I	II	III	I	II	III
p25									
Conditioned	-1.33	-1.42	-1.33	-0.05	-0.68	-0.24	0.64	0.37	-0.38
	(3.11)	(3.2)	(2.97)	(3.46)	(3.68)	(3.82)	(4.29)	(4.23)	(4.33)
Rounding		0.09	3.04		2.62	2.82		3.31	4.34
		(2.69)	(3.13)		(3.56)	(4.04)		(4.23)	(4.23)
Recent			7.33**			2.23			7.51
			(3.03)			(3.97)			(4.57)
p50									
Conditioned	-3.29	-2.66	-3.11	1.39	1.63	2.35	-2.82	-4.00	-3.33
	(2.77)	(2.88)	(2.9)	(3.18)	(3.23)	(2.96)	(4.11)	(4.15)	(4.21)
Rounding		-1.76	0.51		4.38	3.44		2.00	2.34
		(3.17)	(3.52)		(3.5)	(3.65)		(3.22)	(3.53)
Recent			4.31			-3.28			1.95
			(3.4)			(3.73)			(4.53)
p75									
Conditioned	-3.80	-3.71	-4.48	0.89	1.17	2.23	-3.13	-3.38	-0.62
	(3.56)	(3.52)	(3.42)	(3.79)	(3.87)	(4.13)	(4.29)	(4.52)	(4.1)
Rounding		-0.98	0.40		3.53	2.42		0.86	2.52
		(3.44)	(3.2)		(4.01)	(4.1)		(4.5)	(4.72)
Recent			5.43*			-2.69			8.38*
			(3.18)			(4.28)			(4.34)
N	1,817	1,817	1,817	1,529	1,529	1,529	1,802	1,802	1,802

* p<.10, **p<.05, *** p<.01. Bootstrapped standard errors are shown in parentheses (1,500 reps). Shown are effects of conditioning on quantiles, respondent age and education are included in models but not shown here.
†Wave 6 data are unweighted, whereas all other data are longitudinally weighted to correct for attrition.

Table 12. Quantile-regression of conditioning, rounding and recent weighing on men's self-reported weight.

	Wave 2			Wave 4			Wave 6†		
	I	II	III	I	II	III	I	II	III
p25									
Conditioned	-1.23	-1.18	-1.51	-1.07	-2.14	-1.47	-0.09	0	0.64
	(3.09)	(3.09)	(3.05)	(3.02)	(2.92)	(3.03)	(3.06)	(3.4)	(3.37)
Rounding		4.66	3.61		2.63	2.82		2.4	4.85
		(2.97)	(2.97)		(3.21)	(3.24)		(3.07)	(3.36)
Recent			0.82			3.82			5.92**
			(2.6)			(3.42)			(2.59)
p50									
Conditioned	0.27	-0.32	-0.63	-0.22	-0.39	-0.25	2.83	2.19	3.55
	(2.45)	(2.12)	(2.28)	(2.91)	(2.66)	(2.66)	(2.99)	(3.02)	(2.85)
Rounding		5.44**	5.53**		4.53*	4.4		3.88	3.64
		(2.26)	(2.41)		(2.61)	(2.95)		(2.78)	(2.66)
Recent			-0.29			-0.1			4.67*
			(2.27)			(2.88)			(2.71)
p75									
Conditioned	5.27**	7.07**	7.04**	1.52	2.82	3.1	6.13	5.97	5.2
	(2.97)	(3.05)	(3.17)	(3.94)	-(4.2)	-(4.)	(5.34)	(5.12)	(5.34)
Rounding		5.35	5.26		6.33	6.60*		0.16	0.25
		(3.28)	(3.24)		(3.91)	(3.97)		(4.47)	-(4.7)
Recent			-0.78			0.03			3.22
			(2.58)			(3.46)			(5.02)
N	1,817	1,817	1,817	1,529	1,529	1,529	1,802	1,802	1,802

* p<.10, **p<.05, *** p<.01. Bootstrapped standard errors are shown in parentheses (1,500 reps). Shown are effects of conditioning on quartiles, respondent age and education are included in models but not shown here.
†Wave 6 data are unweighted, whereas all other data are longitudinally weighted to correct for attrition.

Table 13 shows estimates of panel conditioning effects on quartiles of self-reported height as well as calculated relative weight, the "Body-Mass Index" (BMI). Results for men's self-reported height are consistent with H1 across all waves: Conditioned men are significantly more likely to report a taller height than unconditioned men in the tallest quartile of height. There is no effect of conditioning on men's calculated body-mass. Moreover, there is no effect of conditioning on women's self-reported height or on women's calculated body mass.

Table 13. Quantile regression of height and body-mass index on panel conditioning, Waves 2, 4 and 6

	Wave 2		Wave 4		Wave 6†	
Men	Inches	BMI	Inches	BMI	Inches	BMI
q25						
Conditioned	0.16	-0.41	0.22	-0.08	0.35	-0.69
	(0.27)	(0.4)	(0.34)	(0.41)	(0.4)	(0.58)
q50						
Conditioned	0.38	-0.54	0.22	-0.07	0.56	-0.27
	(0.28)	(0.36)	(0.35)	(0.48)	(0.45)	(0.47)
q75						
Conditioned	0.62***	-0.42	0.75**	0.36	0.78**	-0.78
	(0.24)	(0.43)	(0.32)	(0.57)	(0.34)	(0.55)
N	1,817	1,817	1,529	1,529	1,802	1,802
Women	Inches	BMI	Inches	BMI	Inches	BMI
q25						
Conditioned	0.16	-0.11	0.21	-0.42	0.64*	-0.11
	(0.23)	(0.39)	(0.21)	(0.41)	(0.36)	(0.41)
q50						
Conditioned	0.13	-0.37	0.16	-0.15	0.11	0.09
	(0.26)	(0.49)	(0.29)	(0.44)	(0.33)	(0.68)
q75						
Conditioned	0.1	0.93	0.17	0.11	0.17	0.03
	(0.23)	(0.67)	(0.33)	(0.63)	(0.31)	(0.99)
N	1,817	1,817	1,529	1,529	1,802	1,802

* p<.10, **p<.05, *** p<.01, Bootstrapped standard errors are shown in parentheses (1,500 reps). Shown are effects of conditioning on quartiles, respondent age and education are included in models but not shown here.
†Wave 6 data are unweighted, whereas all other data are longitudinally weighted to correct for attrition.

Table 14 shows the effects of panel conditioning on other indicators of weight response quality. H2 is not supported for men: Conditioned men are no more or no less likely to non-respond to a question about their weight. Conditioned women, on the other hand, do seem to be influenced by conditioning in their propensity to non-respond though the results are not statistically significant at Wave 2 and Wave 6, but are significant at Wave 4.

Table 14. The effects of panel conditioning on self-reported weight response quality, Waves 2, 4 and 6.

Men		Rounding	Rounding	Weight NR	Weight NR	Recent Weighing
Wave 2	Conditioned	0.02	0.00	0.05	-0.02	-0.19
		(0.2)	(0.21)	(0.61)	(0.62)	(0.16)
	Recent weighing		-1.13***		-1.41*	
			-0.19		-0.76	
	N	678	672	685	679	781
Wave 4	Conditioned	-0.26	-0.29	2.17**	-----†	-0.19
		(0.24)	(0.24)	(1.09)	-----†	(0.25)
	Recent weighing		-1.21***		-----†	
			(0.28)		-----†	
	N	550	548	507	293	616
Wave 6	Conditioned	-0.25	-0.25	0.86	0.73	0.11
		(0.23)	(0.23)	(0.66)	(0.65)	(0.21)
	Recent weighing		-0.59***		-1.57*	
			(0.22)		(0.82)	
	N	596	596	617	616	763

Women		Rounding	Rounding	Weight NR	Weight NR	Recent Weighing
Wave 2	Conditioned	-0.08	-0.05	-0.78	-0.65	0.21
		(0.17)	(0.17)	(0.6)	(0.61)	(0.14)
	Recent weighing		-0.57***		-1.10**	
			(0.17)		(0.52)	
	N	769	766	792	787	918
Wave 4	Conditioned	-0.09	-0.1	-1.22**	-1.51**	-0.14
		(0.22)	(0.22)	(0.57)	(0.7)	(0.19)
	Recent weighing		-0.46**		-0.76	
			(0.22)		(0.57)	
	N	655	654	678	672	746
Wave 6	Conditioned	-0.05	-0.05	-0.13	-0.01	0.05
		(0.2)	(0.2)	(0.36)	(0.38)	(0.18)
	Recent weighing		-0.68***		-1.89***	
			(0.21)		(0.46)	
	N	766	764	833	823	1,007

* p<.10, **p<.05, *** p<.01, Standard errors are shown in parentheses. Respondent age and education are included the model but not shown. Data are weighted to control for attrition, except for Wave 6. † Results cannot be shown because conditioning and recent weighing predict weight response perfectly in Wave 4 for men.

Like H2, H3 is not supported for men. Although not statistically significant, findings for Wave 4 and Wave 6 suggest that conditioned men are less likely to round their answers to the nearest stone or half-stone than unconditioned men. H3 is also not statistically supported for women, though conditioned women are less likely to round their answers as compared to unconditioned women at all waves. It should be noted that recent weighing was controlled in all of these models because knowing one's weight may encourage reporting of a weight that is more accurate. A model predicting recent weighing finds that there is no effect of conditioning for either men or women on recent weighing.

Conclusions.

Given that validation work examining self-reported height and women suggest people tend to report more culturally normative weights, it seems clear that these sorts of survey questions are influenced by social desirability effects. Theoretically, some have argued that continued participation in panel surveys encourages trust and more accurate reporting over the life of the panel. Tested was whether varying the frequency of questions on height and weight encouraged more accurate reporting of these facts premised on the assumption that those most likely to socially desirable respond would be less likely to do so if they experienced panel conditioning. Heavy women were found to report greater weights when conditioned and tall men were found to report higher heights when conditioned. There is some evidence that conditioned women were less likely to round their responses but no clear evidence that non-response was thwarted by conditioning. These results suggest some evidence that panel conditioning may counter social desirability effects in panel surveys; however, these findings are not particularly strong.

e. The impact of changing self-completion formats between paper and computer (S.C. Noah Uhrig)

At Wave 3 of the main UKHLS interview, the adult self-completion instrument was administered using Computer Assisted Self Completion (CASI). This represented a permanent shift from a paper self-completion instrument and was instituted to reduce what was perceived to be unacceptably high levels of unit non-response on the paper instrument at

Waves 1 and 2. An IP experiment was created to examine the effects of shifting the design format of the self-completion instrument. The experimental design covered three waves of data collection (IP4-IP6). Table 15 contains a schematic design of the three-wave experiment. Early results from the first wave of this experiment were discussed in the IP4 working paper.[7]

Table 15. Experimental allocation to self-completion instruments

Wave 4 (D)	Wave 5 (E)	Wave 6 (F)	Comparison Groups	Frequency
1 – CASI	1 – CASI	1 – CASI	A	129
	2 – Paper	1 – CASI	B	123
		2 – Paper	C	108
2 – Paper	1 – CASI	1 – CASI	D	64
		2 – Paper	E	70
	2 – Paper	2 – Paper	F	254

At Waves 5 and 6 where web and face-to-face interviewing were examined in a sequential mixed-mode design, all households were allocated to treatment, however only sample members interviewed face-to-face were subject to this experiment. Thus, allocation to treatment was independent of the mixed-mode experiment, but response to the mixed-mode experiment reduced the overall sample size available for analysis. Also, due to a programming error at Wave 5 around 50 per cent of those eligible to receive the questions in face-to-face CASI mode did not get asked the experimental questions (313 people, based on unedited data). It should be noted that this does not confound the experiment (i.e. no respondents were asked questions in the wrong self-completion format), but this error does reduce its power to detect differences across self-completion formats. For these reasons, only respondents interviewed face-to-face across all three waves who experienced no errors in experimental administration were analysed.

Methods

The key indicators of data quality are the reliability and stability of core self-completion measures. Due to experimentation with much of the core content across the three waves, the

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

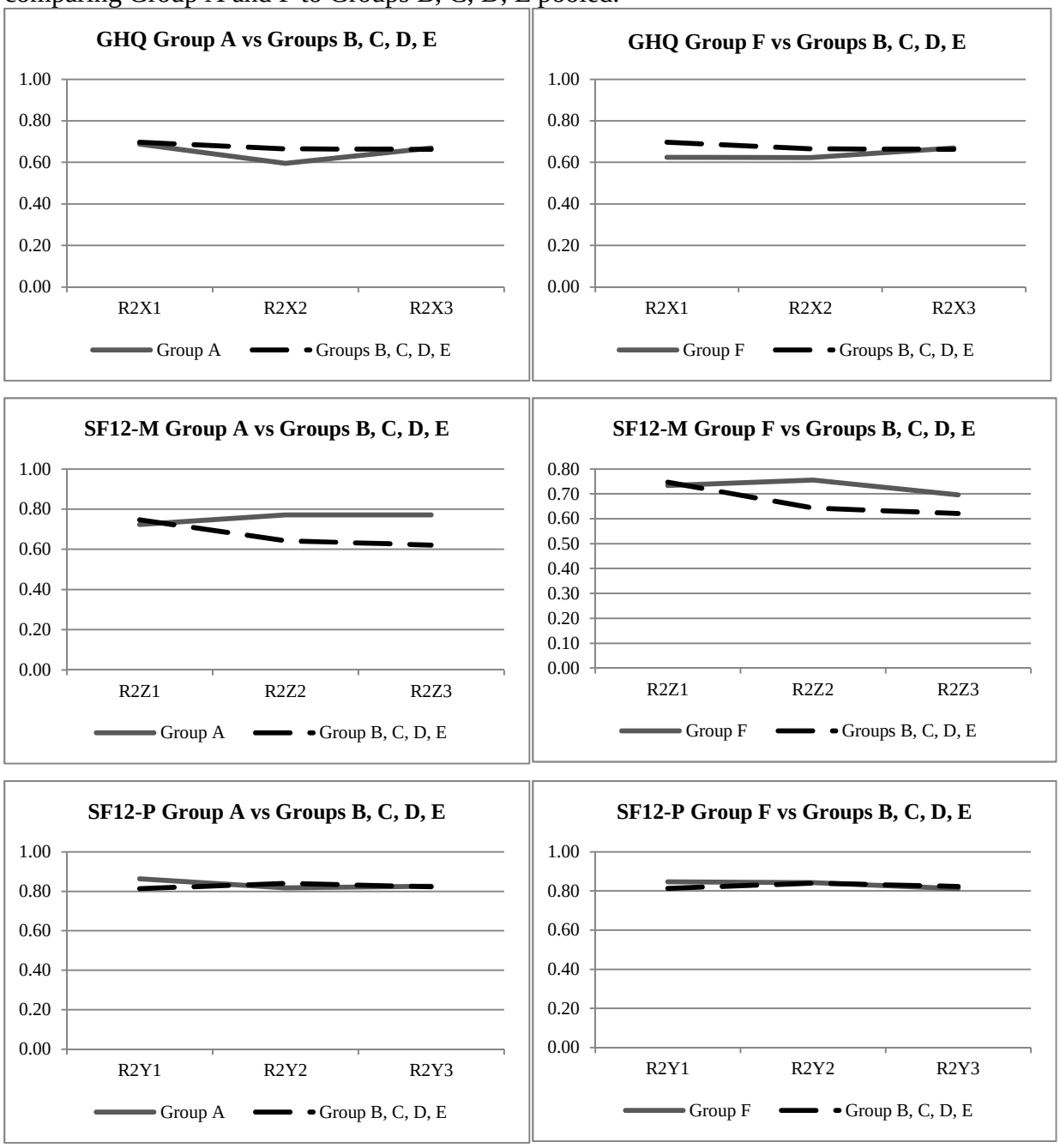
only core items amenable to analysis are the composite SF12 mental (SF12-M) and physical (SF12-P) health scores, and composite GHQ scores. Note that both SF12-M and the GHQ are measures concerned with mental health whereas the SF12-P concerns physical health. We might expect measures of mental health to behave more like subjective survey items whereas physical health measures may take on characteristics of objective survey items.

Quasi-Simplex Models with the three waves of data were used to obtain reliabilities and stabilities for these three core UKHLS measures across each of the six experimental treatment paths. All models were obtained using Bayesian estimation. It should be noted that the confidence intervals suggest non-significant differences across all groups in the results presented, however this is likely to be a consequence of small sample sizes on the experiment. Confidence intervals are only shown on figures where doing so does not reduce readability.

Results

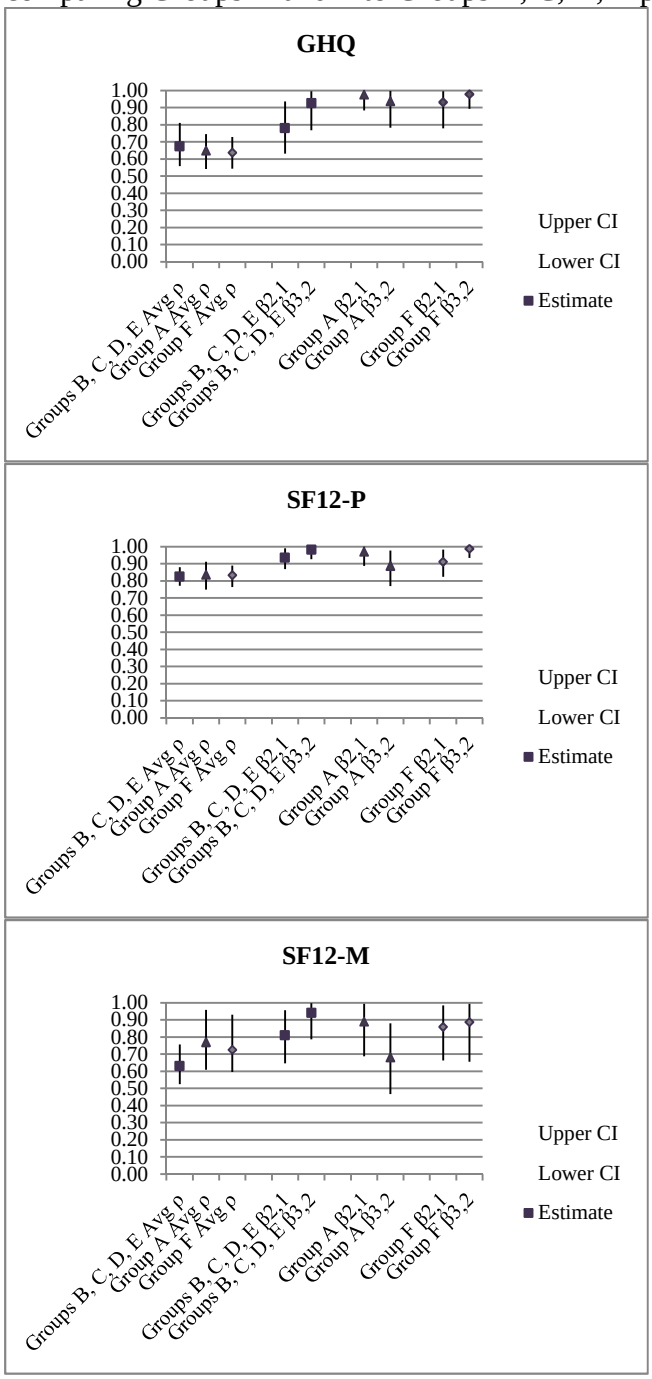
Is there any self-completion format effect at all? A comparison of single-mode versus mode switching with the self-completion instrument would indicate whether there was any mode effect at all. One might expect wave sequential switching of self-completion formats to be associated with lower reliabilities and stabilities given that the visual presentation of items varies across paper and CASI formats. Shown in Figure 2 are sets of charts with three-wave reliability estimates where Groups A (CASI only) and F (Paper only) are compared to a pooling of Groups B, C, D and E (all mixtures of modes across waves). Figure 3 shows average reliabilities and the stability coefficients across these treatments.

Figure 2. Reliabilities at three waves from QSM for the GHQ, SF12-M and SF12-P, comparing Group A and F to Groups B, C, D, E pooled.



.

Figure 3. Average reliabilities and stability coefficients for the GHQ, SF12-M and SF12-P, comparing Groups A and F to Groups B, C, D, E pooled.



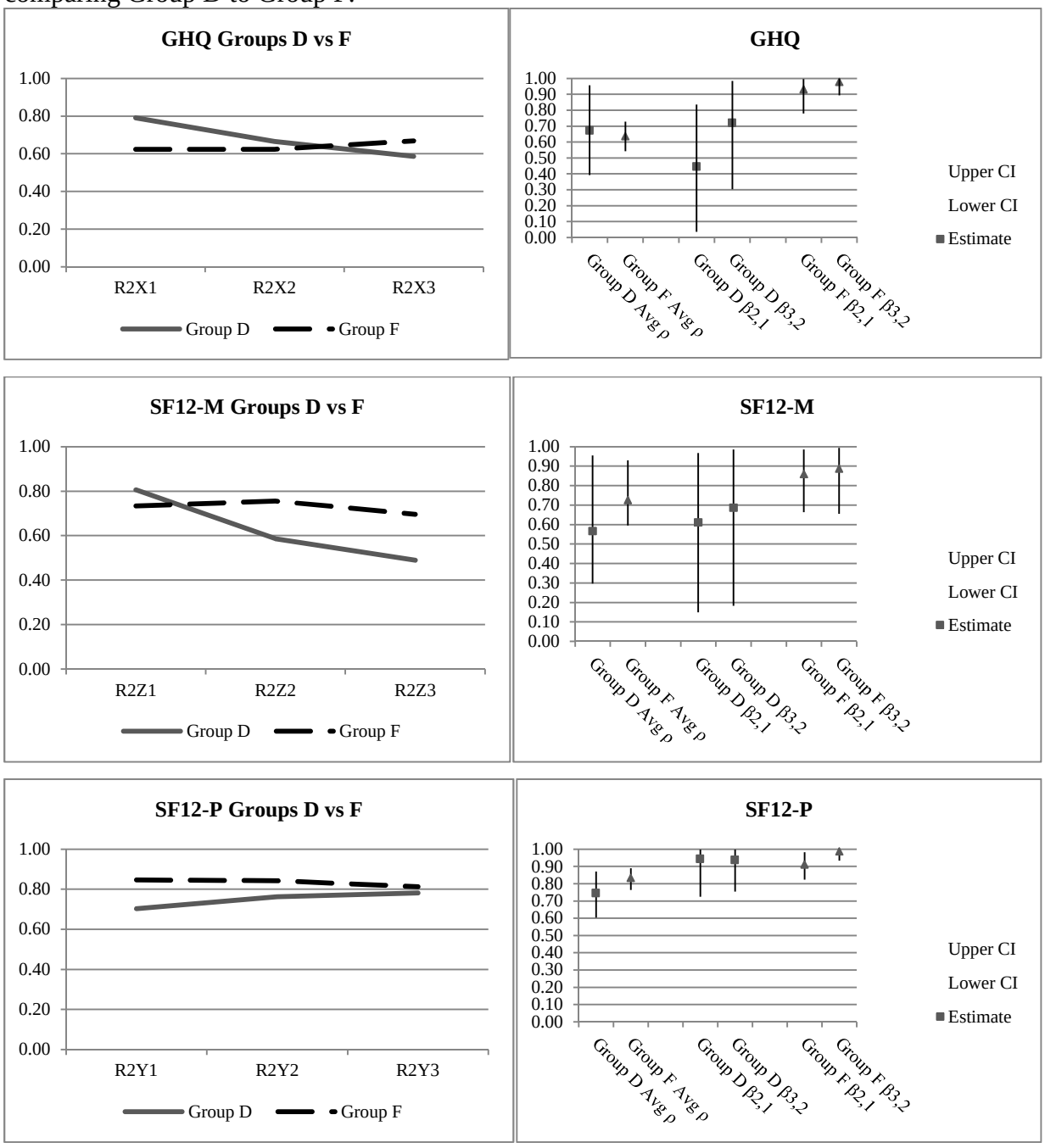
Results suggest that for the GHQ and the SF12-P, there is little effect of format switching across sequential waves as compared to administration by a single self-completion format, either CASI (Group A) or Paper (Group F). However, for the SF12-M, wave consistency in

instrument format is associated with a higher reliability as compared to switching, regardless of the instrument format.

As with the individual wave reliability estimates, average reliabilities across the three measures do not differ for the GHQ and the SF12-P, however instrument consistency is associated with somewhat higher average reliability for the SF12-M as compared to format switching. Stability coefficients are an indicator of how much wave on wave change exists in the true value for each measure. Patterns in stability shifts are consistent across self-completion formats though the magnitudes of stability coefficients differ somewhat across the three measures investigated. We observe a steeper increase in stability coefficients for a wave sequential mixture of questionnaire formats between IP5 to IP6 as compared to IP4 and IP5 in the GHQ and the SF12-M but not for the SF12-P. Similarly, the stability coefficients decline for the SF12-M with CASI only but the declines are lower for the GHQ and the SF12-P under CASI only. Taken together, these results tentatively imply format consistency yields better data than wave sequential format switching for the three measures analysed. There is greater measurement consistency, however, in the more objective SF12-P as compared to the more subjective SF12-M and GHQ.

What is the effect of directional switch from paper to CASI? Comparison of Group D to Group F represents a direct examination of the Wave 3 switch from paper to CASI in the main-stage instrument. Figure 4 shows these results. If the switch to CASI negatively impacts reliability or stability, we would observe a decrease in the reliability coefficients or an appreciable shift in stability coefficients over waves.

Figure 4. Reliability, Average Reliability, and Stability of the GHQ, SF12-M, SF12-P, comparing Group D to Group F.



Results suggest higher and more consistent reliability estimates for Group F (paper only across three waves), both at each wave and in the average reliabilities. Stability coefficients are also generally higher for the paper only treatment in the GHQ and SF12-M but there is little difference between stability coefficients in the SF12-P between experimental treatments.

This implies that consistency with a paper questionnaire yields better data than the period of time around switching from paper to CASI, particularly for subjective measures such as the GHQ and the SF12-M.

What is the effect of reversing the decision to switch to CASI? Anecdotal evidence suggests that using a CASI instrument reduces fieldwork efficiency in a household panel. Interviewers give over their laptops to respondents and the total time interviewers spend in households increases as a function of the amount of time it takes respondents to complete the self-completion instrument. With a paper questionnaire, respondents can get-on with completing the instrument while the interviews with other household members are conducted. Continued interest in improving fieldwork efficiency might encourage a return to using a paper self-completion instrument. Shown in Figure 5 is a comparison of Group D (unidirectional shift to CASI) to Group E (returning to paper after a single wave of CASI). Similarly a comparison of Group D to Group C, shown in Figure 6, represents what would happen after two waves of returning to paper.

Figure 5. Reliability, Average Reliability, Stability for the GHQ, SF12-M and SF12-P, comparing Group D with Group E

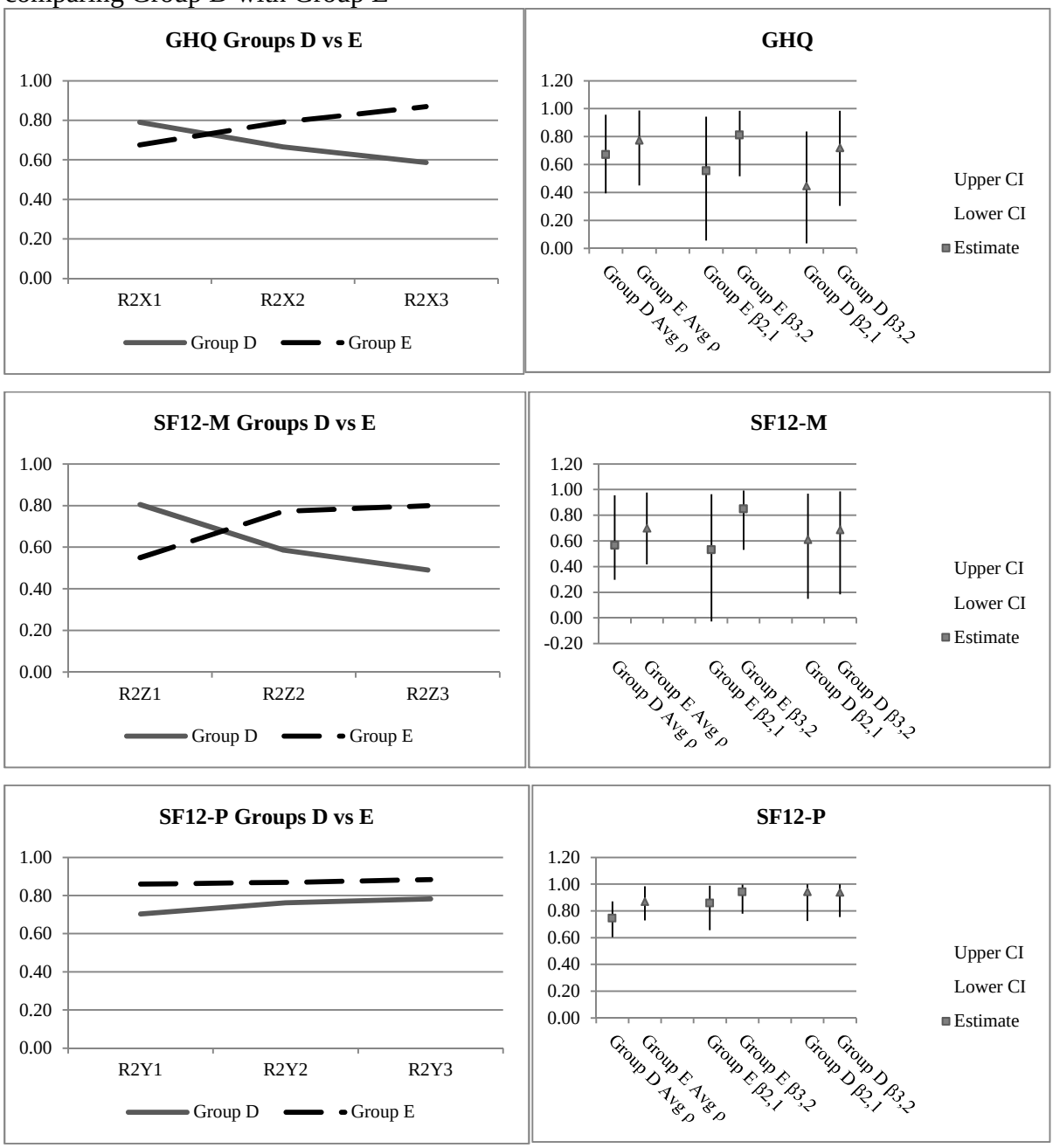
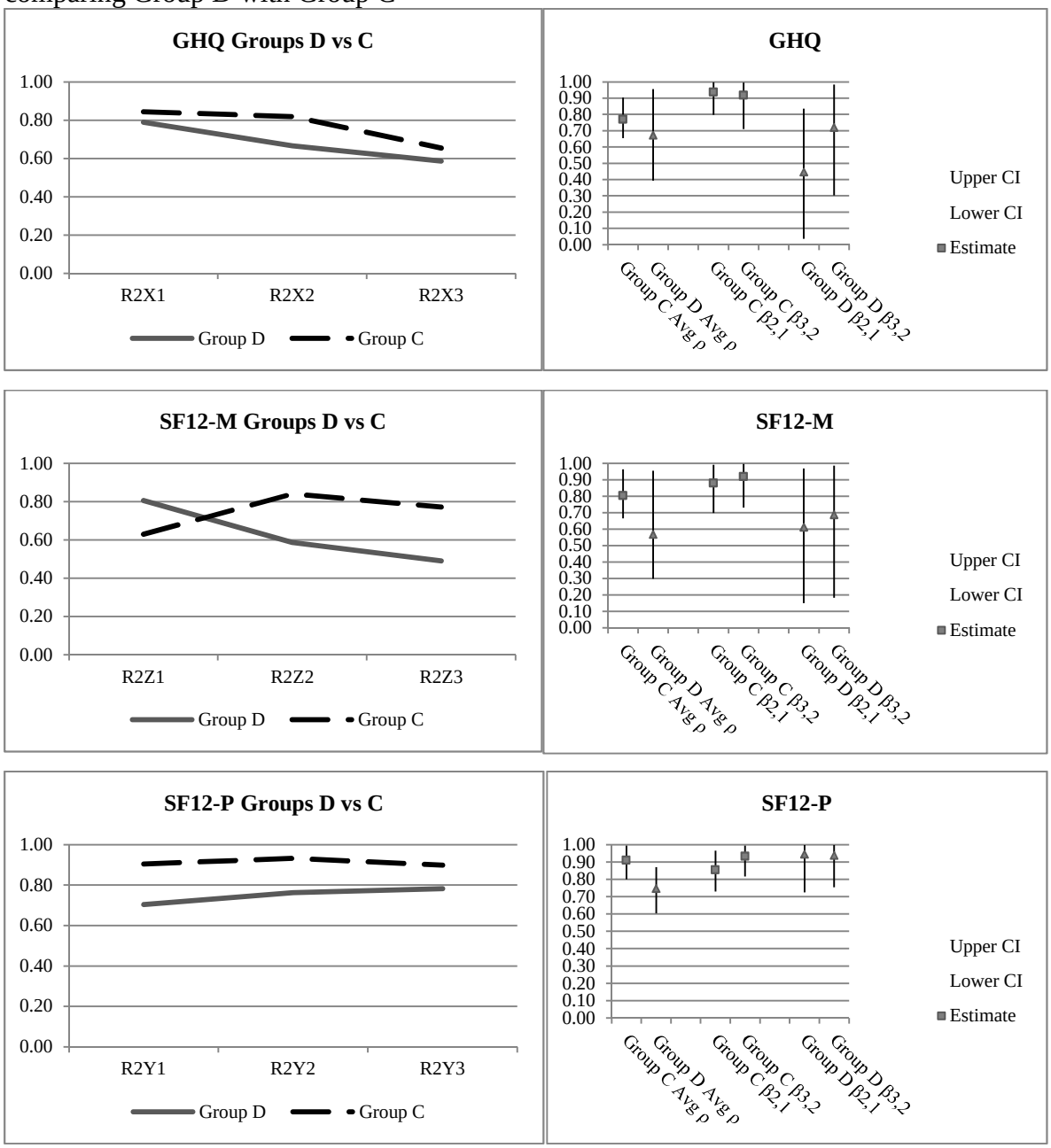


Figure 6. Reliability, Average Reliability, Stability for the GHQ, SF12-M and SF12-P, comparing Group D with Group C



Group D shifts to CASI at IP5 from paper at IP4 and sticks with it through to IP6. Group E reverts to paper at IP6. We find a decline in reliability amongst Group D respondents for the GHQ and SF12-M with a corresponding increase in reliability for Group E respondents. Reliabilities appear stable for the SF12-P over time, though Group E is somewhat higher than

Group D. The GHQ and SF12-M both show an increase in stability for Group E with little effect for the SF12-P. Group E stabilities shift only for the GHQ.

Here, the unilateral shift to CASI in Group D is compared to a return to paper from a point of using CASI in Group C. As with the comparison of returning to paper for a single wave (Group D versus Group E shown in Figure), we improvements in reliability over time with the paper questionnaire relative to the CASI instrument for the SF12-M and an overall higher reliabilities for the GHQ and SF12-M with a return to the paper instrument. Considering stability coefficients, the return to paper would seem to have a higher stability for the GHQ and the SF12-M as compared to the shift to CASI with little difference in the SF12-P.

Summary

Taken together these results suggest that the shift to CASI is associated with less stable reliabilities and less stable lag-1 relationships between true values on core GHQ and SF12 measures around the transition as compared to consistency in self-completion format. Comparisons also suggest that returning to a paper instrument improves reliability for the most part.

f. Measuring Partnership Satisfaction with the Division of Housework (Katrin Auspurg, Maria Iacovou, Cheti Nicoletti)

In all modern societies, there are gender differences in the allocation of work, with women doing the majority of housework (about 60 to 70 percent); women do more housework even in couples where both partners have full-time jobs.

Explanations proposed for the persistence of gendered work arrangements include gender norms and identities; gendered preferences; and gains from specialization arising because men have higher earning power than their female partners. However, most research based on household survey data cannot distinguish fully between these factors, because we observe very few couples where women have higher earning power than their male partners; we therefore do not know what sort of work arrangements would be in place if men and women had comparable earning power.

Our experiment was designed to overcome this problem. We generated a battery of hypothetical scenarios ("vignettes") which varied over five dimensions: the amount of paid work, the relative earning power of male and female partners; the presence and age of children; the division of paid und unpaid work; and whether the couple has paid help with the housework. The exact wording of all vignettes may be found in the IP questionnaire documents; a sample scenario reads as follows:

> **Imagine that you are married or cohabiting, you and your partner both have full time jobs, and your partner has an hourly pay which is twice as much as yours**. **You have no children**, **your partner does all of the housework while you do none of it**, **but you employ somebody to help with the housework one morning per week**.
> **How satisfied would you say you are with the sharing of the housework?**
> Respondents were asked to reply on a seven-point scale, from 1 (completely dissatisfied" to 7 "completely satisfied").

For each respondent, three vignettes were selected at random from the full battery. Respondents were asked to imagine themselves in these hypothetical scenarios, and to rate how satisfied they would be with the allocation of household work in each scenario. This experiment was first run in IP5; it was repeated in IP6, with all respondents being asked to rate *exactly the same set of vignettes* that they had already rated in IP5.

The aims of the repeated experiment were (1) to test the stability over time of satisfaction ratings, (2) to analyse possible reasons for any changes in these ratings; and (3) to contribute to research on the validation of factorial surveys and anchoring vignettes;[8] only a few studies have used these methods in the context of longitudinal surveys.

[8] There is concern on the comparability of reported satisfaction levels between individuals when using common item questions: different individuals might use different reference standards for rating their satisfaction. Working with repeated measurements on highly standardized scenarios like the "anchoring vignettes methodology" used in our experiments may address these problems of comparability.

Results: data quality and overall response patterns

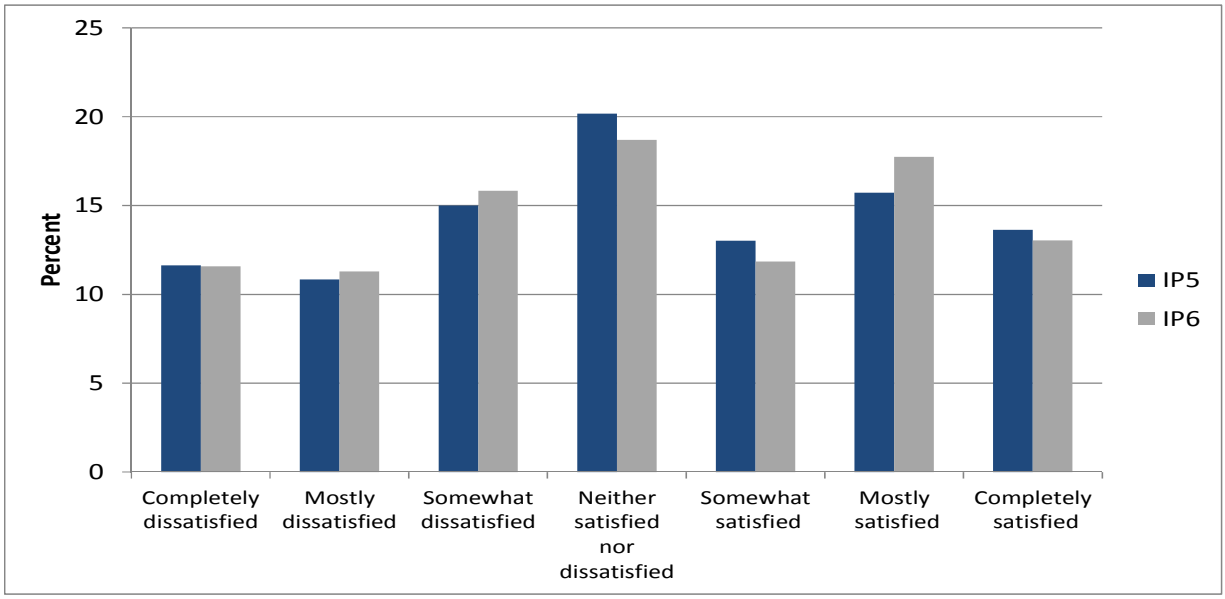
1,310 respondents participated in both the IP5 and IP6 experiments. As all respondents rated multiple scenarios, there are more cases than respondents, with pairs of repeated measures for 3,618 scenarios.

For experimental designs it is crucial that (1) the questions allocated to each respondent are uncorrelated with respondent characteristics; (2) the factors varying between questions are not cross-correlated; and (3) all levels of the experimental factors occur with about the same frequency. Analyses showed all these target criteria to be met in both IP5 and IP6. All correlations between experimental factors and respondent characteristics (age, labour market and family status, etc.) were below r = 0.1, and there was no significant correlation between the different experimental factors.

93.8% of IP6 respondents provided valid ratings for all three questions (IP5: 92.7%); 3.2% refused or answered “don’t know” to all three vignette ratings (IP5: 4.0%); the remainder gave valid answers to one or two questions. The proportion of valid responses over all vignettes was 95.3% (IP5: 94.4%).

In both IP5 and IP6, a high proportion of respondents used the same response category for all three questions (32.3% for both IP5 and IP6); these respondents were significantly more likely than others to give the middle response category (28.0% versus 14% in IP6; 31.0% versus 14.9% in IP5). Figure 7 shows the distribution of vignette ratings by panel wave. The distribution of vignette ratings is very stable over time (t-test for mean differences: $t = 0.005$; $p = 0.962$; $n = 9{,}622$).

Figure 7: Ratings of vignettes by panel wave



Note: IP5: *n*=4,556 vignette ratings (*n*=1,544 respondents); IP6: 5,066 vignette ratings (*n*=1,716 respondents).

Results: stability of responses across panel waves

In this section we consider changes in individuals' satisfaction ratings with the same vignette. In 28% of cases, ratings did not change. In a further 30% of cases, the rating increased or decreased by 1, and in 19% of cases it increased or decreased by 2. Thus, a large majority of the ratings changed not at all, or by a relatively modest amount. However, in almost one quarter of cases (23.2%) the ratings changed by 3 points or more.

Since respondents in both waves were presented with exactly the same scenarios, these changes indicate either measurement problems (low reliability caused by problems such as respondents' fatigue or misunderstanding of the task), or real changes in preferences that occurred over time.

Initial analyses revealed that changes in vignette ratings are not correlated with the following:

- Respondents' age (older respondents may have more problems with complex survey tasks).

- The “spread” of ratings which respondents used for the different scenarios (a low spread may represent misunderstanding of the task or low willingness to provide considered answers - so called “satisficing”).
- Respondents’ partnership status (people living in a partnership at time of the survey may find it easier or more difficult to imagine themselves in the different hypothetical scenarios).
- Respondents whose own partnership status changed from year to year (e.g. those who lived with a partnership in IP6 but not in IP5, or vice versa) showed a small tendency to make more changes in ratings, but this effect was not statistically significant.

By contrast, changes in vignette ratings are strongly and significantly correlated with changes in respondents’ reported satisfaction with their own actual housework arrangements.

These findings suggest that the rather low test-retest reliability of responses is not caused by methodological problems, but rather, that changes in vignette ratings over time are driven by substantive changes in respondents’ preferences.

Other factors predicting changes in vignette ratings may include factors relating to survey completion (such as very short response times) or additional factors relating to changes in preferences (such as further changes in respondents’ actual living conditions). Further work is necessary to investigate these.

As well as their methodological applications, these data also have useful substantive applications, particularly in research on the reasons behind gender-specific work arrangements.

g. Assessing the effects of prenatal hormone exposure on the human life-course (Cara Booker and Sebastian Schnettler)

Background

According to the organizational hypothesis in behavioural endocrinology, early exposure to androgens has permanent, organizational effects on brain and behaviour (Breedlove 2010,

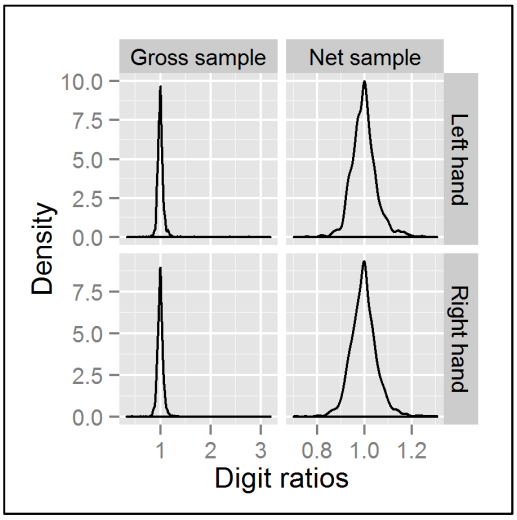
Nelson 2011). These are distinct from activational effects of circulating hormone concentrations. Whereas organizational effects link early environment with behavioural outcomes later in life by eliciting different developmental strategies, activational effects orchestrate behaviour in a more immediate way and in response to changes over the life course. However, recent research in behavioural endocrinology increasingly shows how hormonal effects on behaviour are strongly moderated by social context (Gettler et al. 2011, Taylor 2014). For researchers who thus want to bring together the strengths of social scientific explanations on human behaviour with insights from behavioural endocrinology rely on data that combine hormonal measurements with in-depth socioeconomic information. Although measurement of circulating hormone concentrations is becoming more frequent in surveys, markers for prenatal hormones have not yet been systematically implemented in large-scale, representative surveys with in-depth socioeconomic information that social scientists typically take into consideration. Here, we provide a brief descriptive report on a new data set that combines a measure of prenatal androgen exposure with detailed longitudinal data on socioeconomic background and various life course outcomes. This data set will allow researchers to follow an integrated research perspective on human behaviour, combining insight from behavioural endocrinology with that from the social sciences.

Descriptive Results

The difficulty and costs of direct measurement in embryos and the lag between measurement and later life outcomes has resulted in the need for indirect measurement of prenatal hormones (Breedlove 2010). The length ratio of the second and fourth digits (2D:4D) has been found to be a stable marker of prenatal steroid hormone exposure with high validity (Breedlove 2010, Hönekopp et al. 2007, Manning et al. 1998). As part of IP6 , a 2D:4D module was administered to a representative sample of 2,018 individuals in 1,187 households and the resulting measurements were matched with longitudinal information on various life domains from six waves of the Innovation Panel. From the survey participants, we obtained a total of 1,583 right hand and 1,582 left hand measures on the respective digit ratios. We deleted digit ratios of individuals with issues like broken fingers or severe arthritis on one or both of their hands, as this could have produced invalid digit ratios. Given some extreme values, we further deleted outliers with digit ratios higher or lower than three standard

deviations from the mean (cf. Lippa 2003). Together, this reduced the number of valid cases to 1,468 right hand measurements and 1,490 left hand measurements. The kurtosis and skewness levels of the gross distribution were considerably reduced by these procedures. Kurtosis was reduced from 151.0 and 173.0 to 3.2 and 2.1, respectively, for right and left hand measurements. Skewness was reduced from 8.6 and 9.8 to 0.3 and 0.5, respectively (see Figure 8[9]).

Figure 8. Probability densities before/after removal of outliers and hand injuries.



Overall, the share of missing values corresponds to 27.2% in case of right hand measurements and to 26.1% in case of left hand measurements. The top three reasons for non-participation were un-willingness to participate (n=220), unavailability of a measurement device (n=174), and hand or finger injury (n=143). Interviewers in the face-to-face were equipped with callipers. Thus, unavailability of a measurement device was almost exclusively an issue in the Web-based interviews, leading to a strong association between the share of missing cases and the mode of data collection. The share of individuals without either a left or right hand measurement corresponds to 40.3% of all participants in the Web module, but only to 11.4% of all participants in the face-to-face module. If every respondent who was willing to participate in the 2D:4D module but was lacking a measurement device had had a

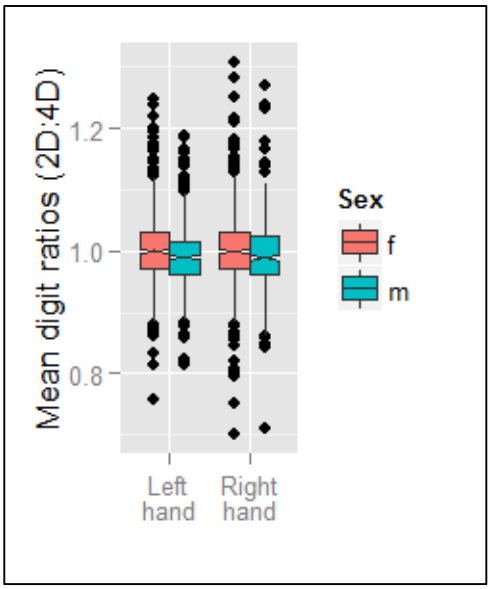
[9] All figures were produced with the *ggplot2* package in R.

ruler available, this would have reduced the missing rate to 19.6%, a level much closer to the missing rate in the face-to-face module.

Consistent with previous research, we find that, on average, male digit ratios are smaller ($M_{m,\ right}$ = 0.993; $M_{m,\ left}$ = 0.993) than female digit ratios ($M_{f,\ right}$ = 1.001; $M_{f,\ left}$ =1.003). These difference are statistically significant, but effect sizes are (Cohen's d) rather small (d_{right} = 0.15, df = 458, t = 2.83, p = 0.005; d_{left} = 0.19, df = 468, t = 3.65, p < .001). The larger sex difference in left hand ratios as compared to right hand ratios is in contradiction with previous findings (Hönekopp et al. 2010). Although the effect sizes of the sex differences we find are smaller than the ones reported in a similarly-sized study, they are within the typical range for 2D:4D studies found in a meta-analyses on digit ratio research (Lippa 2003, Hönekopp et al. 2010). One reason why effect sizes could be relatively small compared to other studies is that we have results from a representative national sample, whereas other results are from special population subgroups or results from convenience samples.

Mean digit ratios between left and right hand measures differ significantly between modes of data collection. Respondents in Web interviews have slightly smaller ratios (M_{Web} = 1.000) than respondents in face-to-face (M_{f2f} = 0.994) interviews (d = 0.14, df = 822, t= 2.43, p=0.015). Looking at this difference separately for males and females reveals that this difference is largely driven by the latter (women: d =0.17, df = 454, t = 2.08, p = 0.038; men: d=0.12, df =0.369, t =1.43, p =0.155) (see Figure. 9). Given preliminary evidence for an association between digit ratios and socioeconomic status (Coates et al. 2009, Hell and Päßler 2011, Putz et al. 2004, but also see Voracek et al. 2010a), one may speculate that this difference may be a result of differential selection by occupational or socioeconomic status into Web-based and face-to-face samples. We will explore this issue in our subsequent analyses.

Figure 9. Notched boxplots of mean digit ratios by hand and sex.



Next Steps

As is known from a large and growing number of lab experiments, 2D:4D is associated with a variety of health-related, physiological, personality and behavioural traits (Grimbos et al. 2010, Hönekopp and Watson 2011, Hönekopp et al. 2006, Voracek and Loibl 2009, Voracek, Pum and Dressler 2010, Voaracek, Tran and Dressler 2010, Voracek et al. 2011) with important between- and within-gender differences (Grimbos et al. 2010, Hönekopp et al. 2010, Hönekopp and Watson 2011, Voracek et al. 2011). There are outcome implications of these traits in various life-course domains such as occupational careers, partnership, reproduction, and adolescent risk behaviour. Existing studies often have few participants and limited information on later-life outcomes. Therefore, they don't allow assessing the external validity of perinatal hormone exposure for life-course outcomes and the hormone and social context interplay throughout the life-course. The data set presented here allows for analysis with higher external validity. Examples of some of our next steps are to:

- estimate the degree of 2D:4D variation in a national population, assess between- and within-gender and regional variation, and additional subgroup descriptive analyses;
- evaluate the feasibility of 2D:4D data collection for wider implementation in cross-national surveys (consent rate, reliability, Web-based vs. face-to-face vs. telephone measurement);
- assess the predictive validity of 2D:4D for life-course outcomes (e.g. educational/ occupational choice, partnership/reproduction, risk for divorce, adolescent risk-taking, cf. Hönekopp et al. 2006, Hönekopp et al. 2010, Shanahan et al. 2003, James 2001).

- study the interaction of social context and hormones on these life-course outcomes;
- assess how single, but possibly interdependent traits (e.g. occupational choice, personality, or risk-taking) play out in synchrony over the life course (Do small associations of 2D:4D on single traits translate into bigger effects in the life course as a whole?).

h. Targeted advance/invitation letters (Peter Lynn)

Lynn (2014) suggests that longitudinal surveys provide ample opportunities to target various features of design or implementation to subgroups in ways that might enhance the likelihood of participation. One such feature is the letter that is mailed to all sample members at the start of each wave of fieldwork. At IP6 an experiment was implemented to test the effectiveness of targeting this letter to particular sample subgroups. A random half of the sample members received a standard letter, designed to have broad appeal. This is the approach that is taken on most surveys and which had been taken at each previous wave of the IP. Sample members in the other random half of the sample received one of six versions of the letter, depending on their socio-demographic characteristics. Much of the content of the letter was the same in each version (e.g. paragraphs about how to complete online, incentives, preparing information in advance, and the voluntary nature of participation), but the opening paragraph was designed to appeal particularly to people with the relevant characteristics. The six versions of the opening paragraph are presented in Table 16.

Table 16: Wording variations in the advance/invitation letter

First paragraph of the letter (for previous-wave respondents):	Thank you so much for helping with the Understanding Society survey last year. The survey helps researchers and policy makers understand the changes in the needs of the country across diverse subjects like <*text*> – and because your information was so valuable, we'd like to hear from you again.
Letter version	<*text*>
Employment-busy	your work-life balance, your position in your employment and your retirement
With dependent children under 15	the provision of child care, schooling and education
Aged 16 to 29	the impact of the economic climate on employment prospects and the influence of mobile technology on life
London & south-east	the cost of living and the provision of schools, housing and public transport
Pensionable age	the provision of social care and the cost of energy and fuel

The second sentence of the standard version of the letter read simply, "The survey helps researchers and policy makers understand the changes in the needs of the country – and because your information was so valuable, we'd like to hear from you again."

At IP6, for one-third of sample households data collection used a single-mode face-to-face design. For these cases the letter sent at the start of field work was therefore an *advance letter*, as it arrived in advance of the actual request for an interview, preparing the sample member for the imminent visit of an interviewer. For the other two-thirds of households, the initial request was to complete the survey online. Only if the survey was not completed online within two weeks did face-to-face attempts then commence (see section 3 of this paper). For these cases, the letter was therefore an *invitation letter*, as it included an invitation and instructions to complete the survey online. In the initial analyses presented here, no distinction is made between the single-mode and mixed-mode samples.

In the half-sample designated for targeted treatment, allocation to a letter version proceeded on a priority basis. The priority order of the five targeted letters was: 1. Employment-busy (usual hours at least 39 per week, or usual hours at least 30 per week with a usual commute of at least 60 hours per week), 2. Has at least one dependent child under 15, 3. Aged 16 to 29, 4. Living in London or the south east (at the time of their most recent interview), 5. Pensionable age. Thus, for example, an employment-busy person aged 16-29 living in London would have received the employment-busy letter, whereas any other person aged 16-29 living in London would have received the aged 16-29 letter. Any person who did not meet any of the five targeting criteria received the standard letter.

The aim of the experiment was to assess the overall effect of the targeting strategy on response rates and response speed, and the effect separately for each of the targeted subgroups. Initial estimates of the effects on response rates are presented here.

Across the issued sample as a whole, the individual response rate at IP6 was 91.0% with the standard letter and 91.2% with the targeting strategy (p = 0.88; n=2,323). There is therefore no evidence of an overall effect on response rate. However, the effect appears to be heterogeneous across sample subgroups (Table 17). Splitting the sample into ten groups, defined by the five targeting groups outlined above and whether or not the sample member responded at IP5, we find a significant effect of the targeted letter for two of the ten groups.

For previous-wave respondents in London and the south east, the targeted letter increased response from 91.9% to 99.1% (P=0.01; n=216) and for previous-wave non-respondents aged 16 to 29 the targeted letter increased response from 54.8% to 82.1% (P=0.01; n=70). The reasons for the specific effects amongst these groups warrant further investigation. The results suggest, however, that targeted letters do indeed have the potential to bring about positive results in terms of response rates.

Table 17. Response rate differences between standard and targeted letters, by previous wave response status and demographic subgroup

	Standard	Targeted	*P*	*N*
IP5 Respondents				
All	97.0	96.7	0.75	1789
Employment-busy	96.3	96.2	0.96	294
With children under 15	99.1	99.2	0.99	233
Aged 16 to 29	97.6	92.5	0.12	176
London & south-east	91.9	99.1	0.01	216
Pensionable age	99.5	98.9	0.48	382
IP5 Non-respondents				
All	70.8	73.0	0.57	534
Employment-busy	84.9	79.4	0.56	67
With children under 15	80.0	90.5	0.34	41
Aged 16 to 29	54.8	82.1	0.01	70
London & south-east	71.4	67.5	0.71	75
Pensionable age	78.3	79.0	0.96	42

i. Mode preferences (Olena Kaminska and Peter Lynn)

Expressed mode preferences in a previous wave has been found to be predictive of mode choice for participation in the following wave (Kaminska and Lynn 2013). Yet, in a long-term panel where a mixed-mode option is planned to continue for a number of waves the question arises whether we need to ask for mode preference in each wave or whether asking for such preference once is enough. In this section we provide preliminary exploration on 1) whether expressed mode preference changes from wave to wave; and 2) whether measuring mode preference repeatedly helps in predicting mode choice in future waves.

The experiment has been carried out over three waves of Innovation Panel: IP4, IP5 and IP6. In all three waves we asked respondents about their preference for mode of participation in the following wave. Five questions on mode preference were asked, but here we focus on one question which asks to choose among four modes: face-to-face, telephone, paper and pencil self-completion and web. Respondents could volunteer ‘no preference’ answer as well. Importantly, IP4 was fielded entirely in face-to-face mode. In IP5 a random 1/3 of households continued with face-to-face (F2F) mode protocol as before, while the other random 2/3 of

households were assigned to a mixed mode condition (MM). In the mixed-mode condition, respondents were offered web mode first, and nonrespondents were followed by face-to-face mode. IP6 followed the same protocol as IP5 (see section 3 for more description). The analysis presented here is restricted to those who participated in all three of these waves.

We find that stability of expressed mode preference between consecutive waves is not as high as may be expected: overall around 60% of people suggested the same mode preference in IP5 as in IP4 (See Table 18). This proportion was higher (73%) between IP5 and IP6. One may suggest that this pattern may be observed due to mode effect (in IP5 some people were asked the questions in web mode). Yet, the pattern remains present if we restrict the observations to F2F condition (those who were never offered web mode) – 65% for IP4-IP5 and 73% for IP5-IP6.

Table 18. Stability of mode preference: proportion of respondents preferring the same mode, by mode assignment group

Sample	Total			F2F			MM		
Stability	Total	F2F	Web	Total	F2F	Web	Total	F2F	Web
IP4-IP5	60.0	65.6	72.0	65.4	79.3	57.7	57.0	58.0	79.5
IP5-IP6	73.1	80.2	81.2	72.9	86.5	65.0	72.9	74.9	85.7

Note: stability represents overall, and then percent stable preferences for a given mode, across sample divisions

Next, focusing on the stability of preference for face-to-face and web surveys (the two modes actually offered), as for all modes generally, stability is higher between the fifth and sixth waves than between IP4 and IP5. Face-to-face mode preference is also more stable in the F2F condition than in the MM condition, while the reverse is true for web mode preference. The web preference is more stable in the MM condition than in the F2F condition. There is some evidence that with time mode preference may stabilize, but we have too little information to conclude it with confidence.

The next question of interest is whether knowing how expressed mode preference changes helps in predicting future mode of participation. Here we restrict our analysis to respondents who participated in all three waves and were assigned to the MM group. Thus, these respondents could participate via either web or face-to-face modes. Here we are interested in predicting IP6 mode participation. First, we explore whether knowledge of mode of participation in IP5 is predictive of IP6 mode of participation (Table 2). We find that it is

highly predictive for IP5 web participants: 92% of those who participated in web last time participated in web again. But it is less predictive of mode of participation among those who participated via face-to-face in IP5: around 28% of them switched to web in IP6. This is interesting and a positive trend since respondents are switching to a mode that is less expensive to administer. It would therefore be useful to be able to predict the propensity to make this switch.

If in addition to previous wave mode participation knowledge we also have the mode preference question – is this helpful in mode assignment for future waves? We explore this for mode preference expressed at IP4 and IP5 separately (no preference includes those expressing no preference in IP4 or IP5). Overall, we find that expressed mode preference does not improve prediction among people who participated in Web mode before (see Table 19). While those who switched later to F2F expressed slightly higher F2F preference – the overwhelming majority in this group still participated in web in IP6. So, regardless of mode preference those who participated in web mode are most likely to respond in web mode again.

Table 19. IP6 Mode participation by IP5 mode participation and IP4 expressed preference

IP5 Mode	F2F				Web			
IP4 Preference	F2F	Web	Other	No Pref.	F2F	Web	Other	No Pref.
IP6-F2F	76.1	50.0	64.5	71.8	11.1	2.7	5.7	7.7
IP6-Web	23.9	50.0	35.5	28.3	88.9	97.3	94.3	92.3
Total	352	76	44	485	252	159	73	507

A different picture is observed among those who participated in F2F mode in IP5: among those who expressed preference for web mode the number of switches to web mode is much higher than among those with other preferences, and the prediction improves when we use IP5 mode preference (63% have switched to web) in comparison to IP4 mode preference (50% have switched to web) (see Table 20). This indicates that at least for those who participated in face-to-face mode repeating the question on mode preference may be useful as it improves prediction of the future potential switchers to web mode.

Table 20. IP6 Mode participation by IP5 mode participation and IP4 expressed preference

IP5 Mode	F2F				Web			
IP5 Preference	F2F	Web	Other	No Pref.	F2F	Web	Other	No Pref.
IP6-F2F	76.9	37.0	68.3	71.8	15.6	5.8	13.3	7.7
IP6-Web	23.1	63.0	31.7	28.3	84.4	94.2	86.7	92.3
Total	373	46	60	485	45	380	75	507

j. Data quality when switching from face to face to web mode in a panel survey (Nick Allum and Fred Conrad)

Face-to-face (FTF) interviews produce population estimates that are widely regarded as the 'gold standard' in social research. Response rates tend to be higher with face-to-face interviews than other modes and face-to-face interviewers can exploit both spoken and visual information about the respondent's performance to help assure high quality data. However, face-to-face interviews are very expensive – considerably more costly than telephone and, especially web, data collection. The question is whether the savings produced by these other modes outweighs any reduction in quality.

A key concern about the web survey data quality is difficulty garnering a probability sample because there are no good frames of email addresses for a general population. In a panel survey, it is possible to switch respondents to web mode after initial recruitment via face to face methods, thus mitigating the problem of sample selection and allowing the collection of rich frame information.

Web respondents generally seem more likely to take shortcuts than respondents in interviewer-administered modes (e.g., Heerwegh and Loosevelt, 2008), but this may be exacerbated by switching from FTF to web: by contrast to an interview, self-administration feels particularly "unsupervised" and, without an interviewer to motivate them to be

conscientious, web respondents may take shortcuts and minimize their effort. This raises the more general issue of whether it is possible to maintain the integrity of time-series in which there is a midstream mode switch (FTF to web).

The main objectives for the present study were:

1. To compare the quality of data on a range of objective and subjective variables provided by respondents of several kinds:

a) F2F respondents

b) F2F respondents who have been randomly assigned to be invited to complete the survey on the web, but who did not take up the invitation

c) Web respondents

2. To evaluate the effect of varying the instructions for respondents on how to go about completing the web survey

Evaluating data quality across mode

The key indicator of data quality for the study capitalises on the fact that this is the sixth sweep of a panel, and will be the consistency between responses in the IP6 experiment and 'validation data' provided in previous waves by the same respondents. Specifically, we selected a range of subjective and objective variables on health and employment and asked respondents to recall their situation when they were first interviewed, in 2008 in IP1. The agreement of the responses and the 'validation data' collected earlier in FTF interviews yield a measure of data quality that we use to assess the relative performance of respondents across modes. Subjective variables present a trickier problem in respect to 'validation' but previous studies have shown that it is possible for survey respondents to recall past mental states or attitudes reasonably accurately (van der Vaart, 2009).

Enhancing data quality in web mode

An important source of measurement error in web surveys can come from respondents who adopt satisficing strategies. In a web survey, people are freed from the necessity of engaging with a face to face interviewer. Because of this respondents are able to miss out on

answering questions, provide non-differentiated responses, spend insufficient cognitive resources on recalling and providing information, and so forth (Chang and Krosnick 2010). One way of mitigating this problem is try to motivate respondents to be conscientious in the online mode, where the risk of inattentive or uncommitted response behaviour is highest. In this approach, we harness respondents' own motivation to complete the survey carefully. One approach of this type is to ask respondents explicitly to commit to answering as accurately as possible. Conrad and his colleagues (Conrad et al., 2011) were able to reduce speeding and straightlining with this approach. In the present study we ask for commitment in a similar way when respondents are first switched to web data collection to see if committed web respondents perform more like they did as FTF respondents than those who are not committed. Objective 2 therefore requires a randomised split sample design where we systematically vary the instructions given to web respondents in completing the relevant parts of the survey.

Selected preliminary results

The following tables 21-23 show the concordance between original answers and answers to the recall questions, for each of three experimental groups of interest. Table 21 examines the general health question, Table 22 the concordance for pain interfering with work and Table 23 the results from the long standing illness question. The percentages sum to 100 in the columns, which means that the estimates represent the probability of giving the observed answer in 2008 conditional on each type of original response. The tables are subdivided in the following way: the first column shows responses for those assigned to face to face interviews in 2008 and in 2013. The second shows responses for those who completed the survey by web in 2013. The third presents the distribution for those who were randomly assigned to complete the survey on the web. This group therefore contains a mixture of web and F2F respondents, as not all those assigned to web completion actually did so. This can be thought of as the 'intention to treat' group, while the web completers are those that actually received the 'treatment'.

Table 21. Percentage of Wave 1 General Health Recalled, for Face-to-Face, Web Complete and Web Assigned

Percentage of Wave 1 General Health Recalled (a_SF1, F_SF1RECALL) Face -to-Face and Web Complete, Web Assign																
Wave 1		Excellent			Very Good			Good			Fair			Poor		
		F2F Assign	WEB	Web Assign	F2F`Assign	WEB	Web Assign	F2F assign	WEB	Web Assign	F2F assign	WEB	Web Assign	F2F assign	WEB	Web Assign
Recall																
	excellent	53	34	35	18	16	17	12	13	10	2	0	5	0	8	3
	very good	34	50	47	48	55	52	42	43	38	15	3	15	0	0	3
	good	8	12	14	27	20	22	30	34	37	44	58	46	11	17	13
	fair	5	4	4	7	8	8	13	8	12	20	29	25	44	33	45
	poor	0	0	0	0	1	1	4	2	3	19	10	10	44	42	37
Count		(85)	(82)	(139)	(123)	(159)	(236)	(94)	(112)	(225)	(54)	(31)	(89)	(27)	(12)	(38)

Table 22. Percentage of Wave 1 Pain Interferes with Work Recalled, for Face-to-Face, Web Complete and Web Assigned

Percentage Pain Interferes with Work Recalled (a_SF5, f_SF5RECALL) Face -to-Face and Web Assign & Complete, Web Assign																
Wave 1		not at all			a little bit			moderately			quite a bit			extremely		
		F2F	WEB	Web Assign	F2F	WEB	Web Assign	F2F	WEB	Web Assign	F2F	WEB	Web Assign	F2F	WEB	Web Assign
Recall																
	not at all	83	84	83	65	52	58	46	32	35	31	44	46	36	50	53
	a little bit	12	13	12	19	31	24	23	32	30	24	11	21	9	17	5
	moderately	3	3	3	10	11	11	15	26	26	3	28	16	9	0	16
	quite a bit	3	0	1	6	5	5	8	11	9	31	11	13	41	17	5
	extremely	0	0	0	0	2	2	8	0	0	10	6	4	5	17	21
Count		(239)	(269)	(444)	(69)	(65)	(131)	(13)	(19)	(46)	(29)	(18)	(56)	(22)	(6)	(19)

Table 23. Percentage of Wave 1 Long Standing Illness Recalled, for Face-to-Face, Web Complete and Web Assigned

Long-standing Illnesses Recall (a_Health, f_HealthRecall) F2F, Web Assign & Complete, Web Assign							
Wave 1		Yes			No		
		F2F	WEB	Web Assign	F2F	WEB	Web Assign
Recall							
	Yes	64	51	57	10	13	11
	No	34	49	43	90	87	89
Count		(129)	(103)	(228)	(253)	(288)	(495)

Tables 24-26 show selected results for the pre-commitment experiment. The tables contrast web respondents who received either the standard or the enhanced encouragement to answer carefully.

Table 24. Percentage of Wave 1 General Health Recalled, by Encouragement

Wave 1		excellent		very good		good		fair		poor	
		encourage	not encouraged	encourage	not encouraged	encourage	not encouraged	encourage	not encouraged	encourage	not encouraged
Recall											
	excellent	35	43	14	18	12	10	0	4	13	0
	very good	49	41	58	49	44	38	8	15	0	2
	good	11	12	22	24	34	35	38	46	25	11
	fair	5	4	5	8	6	13	31	22	25	47
	poor	0	0	0	1	4	3	23	12	38	40
Count		(37)	(187)	(76)	(282)	(50)	(268)	(13)	(130)	(8)	(57)

Table 25. Percentage of Wave 1 Pain Interferes with Work Recalled, by Encouragement

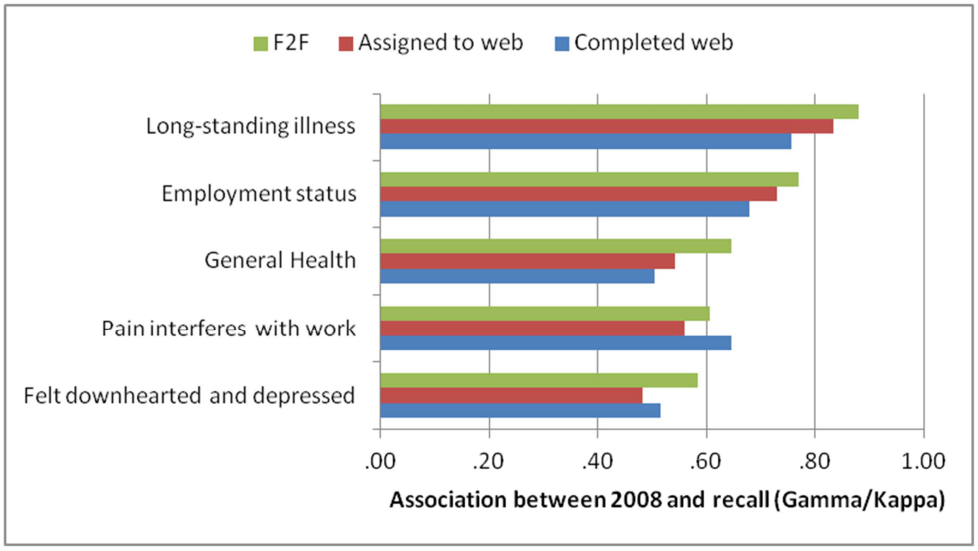
Percentage Web Completed Pain Interfers with work Recall (a_sf5 ,f_sf5Recall) encouraged to take extra time											
Wave 1		not at all		a little bit		moderately		quite a bit		extremely	
		encourage	not encouraged	encourage	not encouraged	encourage	not encouraged	encourage	not encouraged	encourage	not encouraged
Recall											
	not at all	81	84	59	61	42	36	40	41	67	43
	a little bit	16	11	26	21	8	34	0	24	0	8
	moderately	3	3	6	12	42	19	40	10	0	14
	quite a bit	1	2	6	5	8	9	20	19	33	24
	extremely	0	0	3	1	0	2	0	6	0	11
Count		(120)	(563)	(34)	(166)	(12)	(47)	(5)	(80)	(3)	(37)

Table 26. Percentage of Wave 1 Long Standing Illness Recalled, by Encouragement

Percentage Web Completed Long-standing Illness Recall (a_health ,f_healthRecall) encouraged to take extra time					
Wave 1		Yes		No	
		encourage	not encourage	encourage	not encourage
Recall					
	Yes	46	62	11	10
	No	54	38	89	90
Count		(54)	(303)	(130)	(617)

Figure 10 below shows the association between 2008 and 2013 answers for each of the three experimental groups. F2F respondents generally have higher correlations over time between their answers, while those that complete via web are more likely to have the lowest.

Figure 10. Association between 2008 and 2013



Correlations between responses do not tell us how close in absolute terms the concordance is. We calculated a score for each of the health items that are the sum of absolute differences between 2008 and 2013 responses. We then combined these four scores by summing them to create an overall measure of absolute disagreement. Some initial results are shown in the

Figures 11 and 12 below. On this measure, web completers show the lowest disagreement, which is opposite of the trend observed in the pattern of correlations.

Figure 11. Absolute disagreement by F2F and Web completion

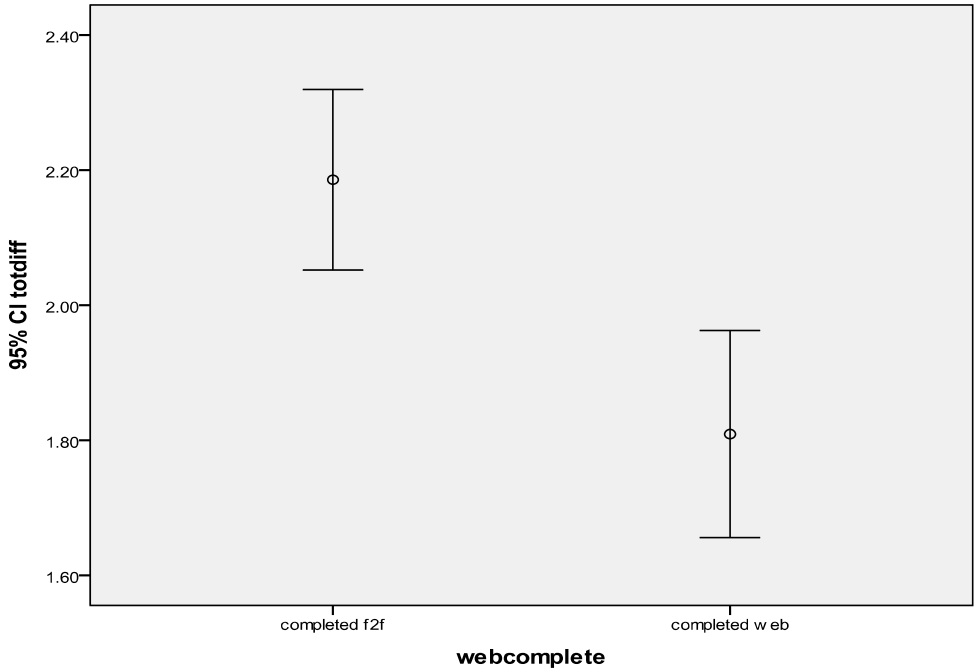


Figure 12. Absolute disagreement by F2F and Web assignment

Overall, from our very preliminary analysis, F2F seems to produce more consistency in recalled answers compared to the originally offered responses, which is what we hypothesised to be the case. However, on an absolute measure of disagreement between original and recall, we find that web respondents have the lowest level of disagreement. Further analyses need to focus on these disparities, as well as adjusting for selection effects in the web completion group.

k. Testing Quick Expenditure Questions (Margaret Blake, Thomas Crossley, Joanna D'Ardenne, Zoe Oldfield, Joachim Winter)

Introduction

Household spending data is important for a wide range of research and policy questions around spending and saving behaviour, and living standards. For example Brewer et al. (2013) demonstrate that UK households with the lowest measured incomes are *not* those with the lowest living standards; household expenditures are a more reliable measure of living standards. Similar findings have been reported for other developed countries.

Household level data on spending has traditionally been collected in national budget surveys, such as the Living Cost and Food Survey in the UK. Such surveys employ methodologies that put a large burden on the respondent, such as expenditure diaries. Consequently, they are not longitudinal and collect limited information in other domains.

Collecting household spending data in longitudinal, multi-faceted studies such as Understanding Society would be particularly valuable. This would allow researchers to study how spending behaviour changes in response to anticipated life-cycle developments (aging, retirement, children leaving home) and to unanticipated shocks (such as job loss or the onset of disability).

In this context, information on *total* household expenditure is desirable. Households are extremely heterogeneous in their spending patterns. This means that spending on particular items can give misleading welfare comparisons between households. Furthermore, food

expenditure is the item most often collected but food is preferentially smoothed in response economic shocks (Browning and Crossley, 2009). This makes it a relatively insensitive measure of household responses to changing circumstances.

There have been a small number of previous attempts to collect total household expenditures with a short sequence of survey questions, including an experiment in the first wave of the Innovation Panel. The results have been mixed.[10] Respondents appear willing and able to provide answers to these questions (item response rates are good) and the data collected contains useful variance. Against this, there is evidence of significant underreporting of expenditure, and cognitive testing has identified problems with the questions.

This IP6 experiment is one part of a larger project to develop quick expenditure questions for use in longitudinal, multi-faceted household surveys.[11] The first part of the project was an intensive process of question development, involving focus groups, consultation with experts and repeated rounds of cognitive testing of alterative designs. The IP6 experiment reported here then tested designs that emerged from that process.

In this experiment we tested a "one-shot" question, in which respondents are asked to report total spending last month against a "breakdown" approach, in which respondents are asked to report spending in a series of categories, and total spending is calculated as the sum of the categories. Past research suggests that a one-shot question will suffer from greater underreporting than a breakdown approach, presumably because asking about individual categories aids in recall. However, the focus groups and cognitive testing we conducted suggested a number of advantages of the one-shot approach. First, the breakdown is more difficult for some respondents and considered more sensitive or intrusive by some respondents. Second, cognitive testing revealed that respondents use different strategies to estimate their spending in the past month. Some add up spending in different categories but others subtract savings from income, for example. The one-shot approach allows respondents to self-select the answer strategy most suitable to the way they think about their budgeting and finances. Focus groups and cognitive testing also identified some key improvements to

[10] See Browning et al. (2014) for a survey.

[11] This project was funded by the Nuffield Foundation. However, the interpretation of the findings expressed here is due to the authors only and does not necessarily represent the views of the Nuffield Foundation.

the one-shot question. These included showing examples and exclusions on show cards; carefully choosing the response unit; and avoiding some problem language. Thus it was important to test these two options after they had been refined by the question development process. The experiment also tested how variants of these two designs work in two different modes: Web and face-two-face (F2F).

Experimental Design

This experiment had a 2 x 2 design. Eligible respondents were randomly allocated to one of two different data collection strategies and to one of two different modes. The table below shows the realized sample

	ONE-SHOT	BREAKDOWN
WEB	231	201
F2F	331	374

The precise one-shot question, in F2F mode was:

"About how much did you [and [NAME OF PARTNER/SPOUSE]] spend on EVERYTHING in the LAST MONTH? Please ***exclude*** *work expenses for which you are reimbursed, money put into savings and repayment of bank loans. Examples of what to include and exclude are shown on this card."*

The web mode was similar, with the exclusions and examples show below the answer box.

The breakdown approach asked for spending in 12 specific categories plus an "other" category. This was followed by reconciliation question which asked:

"So in total, in the last month you [and [NAME OF PARTNER/SPOUSE]] spent [total] pounds. Does that sound right?"

If the respondent answered "*no*", they were then asked:

"How much did you [and [NAME OF PARTNER/SPOUSE]] spend in the last month?"

Two additional elements were added to the one-shot arm of the experiment. First, following up the response strategy heterogeneity uncovered by the cognitive testing, respondents were asked this follow up response-strategy question:

"How did you work out your answer to the spending question?"

Second, in the one-shot approach, we also asked how last month's spending differed from usual:

"Would you say your spending last month was: higher than usual, lower than usual, typical of a usual months spending?"

If higher/lower: *"How much do you [and [NAME OF PARTNER/SPOUSE]] spend on everything in a usual month?"*

Results

Key results from the experiment are presented in Table 26 below.

Table 26. Results from different questions on expenditure, by mode

	Web			**F2F**		
	`One shot'	**`One shot'**	**`Breakdown'**	**`One shot'**	**`One shot'**	**`Breakdown'**
Month:	Last	Usual	Last	Last	Usual	Last
n	195/231	197 / 231	183 / 203	308/331	309/331	370/374
	(84%)	(85%)	(90%)	(93%)	(93%)	(99%)
mean	2260	1646	1807	1312	1219	1801
median	1600	1500	1570	1000	980	1373
Std. dev.	4239	896	1265	1580	1544	1654

Response rates are quite high. As the previous literature suggests, respondents are willing and able to answer questions about their spending.

Because the past literature has found evidence of significant under-reporting in responses to expenditure questions, higher totals are taken as evidence of better data quality. As a benchmark, average monthly household spending in the most recent Living Costs and Food Survey was £2040. This average includes large, complex households (with multiple benefit units) which were excluded in our study, and so is bit too high for our population (which was limited to single benefit-unit households: singles, and couples with dependent children only).

The table above suggests that in F2F mode, the one-shot approach, even as refined by our intensive question development process, suffers from considerably more under-reporting than a breakdown approach. However, this does not seem to be the case in web mode. In web mode, both approaches seem to capture (on average) 80% or more of total spending.

An additional interesting finding from the experiment is that respondents use quite different strategies for answering the one-shot question, depending on the mode in which they are asked. In particular, web-mode seems to facilitate the checking of bank statements and records. It also seems to lead to more heterogeneity in estimation strategies. This is shown in Table 27 below. This may be one explanation for the superior performance of the one-shot equation in web mode.

Table 27. Reporting strategies, by mode

Strategies for answering the spending question (not mutually exclusive)	**Web**	**F2F**
Checked statements	28%	10%
Added up categories	35%	67%
Income minus saving	24%	8%
Recall without checking	21%	15%
Other	5%	5%

Returning to the main results, in web mode, a one-shot question about the last month gives responses which are both higher on average and more variable. This is somewhat puzzling.

Standard infrequency (households make large purchases – durables for example – infrequently, so that in any month, only a fraction of households do so) should increase the variance of responses (relative to "usual" spending) but not the mean. The higher last-month mean here suggests that households were including a fraction or average amount of spending on large-infrequent items in their usual spend. The lower variability of totals from the breakdown approach (particularly in web mode) suggests that this approach delivers better data.

The reconciliation question in the breakdown approach improved data quality considerable, particularly in the web mode. It did so in two ways, one of which we did not anticipate. First, it resulted in the elimination of some outliers and so reduced the variability of reported totals. This was expected. Second, in web (but not F2F) mode, there was quite a bit of item nonresponse to individual categories. In web mode, only 139/203 (69%) of subjects gave a valid response to all 13 categories (including "other"), although 93% answered at least one category.[12] For respondents who answered some, but not all of the categories, we could not know if we had a valid total. The reconciliation question resolved this uncertainty in one of two ways. Many respondents in this group either confirmed that the total of the responses they had given was their total spending for the month, or they were willing to give us a total. Thus reconciliation question raised the fraction of respondents for whom we collected data on total spend to 183/203 (90%). This was a very significant improvement on the initial 69%, and was an unexpected benefit of the reconciliation question. This did not happen in F2F mode where item non-response to spending categories was much lower.

Conclusion

Overall, we conclude that a short-sequence of questions can be used to collect useful expenditure data from households, particularly if delivered in web mode. A breakdown approach with (critically) a reconciliation follow up may deliver the best data, though a one-shot approach delivers quite good data with fewer questions.

[12] The corresponding numbers for F2F mode are 89% and 99%

References

Bailar, Barbara A. 1989. Information Needs, Surveys and Measurement Errors. Pp. 1-24 in *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh. New York: John Wiley.

Borkan, Gary A., David E. Hults, and Robert J. Glynn. 1983. Role of longitudinal change and secular trend in age differences in male body dimensions. *Human Biology* 55:629-641.

Breedlove SM. 2010. Minireview: Organizational Hypothesis: Instances of the Fingerpost. *Endocrinology* 151: 4116–4122. doi:10.1210/en.2010-0041.

Brewer M., B. Etheridge and C.,O'Dea. 2013. Why are households that report the lowest incomes so well-off? *Discuss. Pap. 736*, Econ. Dep., Univ. Essex

Browning, M. and T.F. Crossley. 2009. Shocks, Stocks and Socks: Smoothing Consumption over a Temporary Income Loss *Journal of the European Economic Association,* 7(6):1169-1192.

Browning, M., T.F Crossley and J.K. Winter. 2014. "The Measurement of Household Consumption Expenditures," Annual Review of Economics, (*forthcoming*).

Burton, J. (ed). 2013. *Understanding Society* Innovation Panel Wave 5: Results from Methodological Experiments *Understanding Society Working Paper Series No. 2013 – 06* Institute for Social and Economic Research, University of Essex

Cantor, David. 1989. Substantive Implications of Longitudinal Design Features: The National Crime Survey as a Case Study. Pp. 25-51 in *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh. New York John Wiley.

Coates JM, Gurnell M, Rustichini A. 2009. Second-to-fourth digit ratio predicts success among high-frequency financial traders. *PNAS* 106: 623–628. doi:10.1073/pnas.0810907106.

Conrad, F., Tourangeau. R., Couper, M. & Zhang, C. 2011. Interactive interventions in Web surveys can increase response accuracy. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Phoenix, AZ.

Dekkers, Johanna C., Marieke F. van Wier, Ingrid JM. Hendriksen, Jos WR. Twisk, and Willem van Mechelen. 2008. Accuracy of self-reported body weight, height and waist circumference in a Dutch overweight working population. *BMC Medical Research Methodology* 8:69.

Department of Health 2009. Shaping the Future of Care Together. London: The Stationery Office.

Gettler LT, McDade TW, Feranil AB, Kuzawa CW 2011. Longitudinal evidence that fatherhood decreases testosterone in human males. *PNAS* 108: 16194–16199. doi:10.1073/pnas.1105403108.

Grimbos T, Dawood K, Burriss RP, Zucker KJ, Puts DA 2010. Sexual orientation and the second to fourth finger length ratio: A meta-analysis in men and women. *Behavioral Neuroscience* 124: 278–287

Hell B, Päßler K 2011. Are occupational interests hormonally influenced? The 2D:4D-interest nexus. *Personality and Individual Differences* 51: 376–380.

Heerwegh, D. and Loosveldt, G. 2008. Face-to-face versus web surveying in a high-internet-coverage population. *Public Opinion Quarterly* 72, 836-846

Hönekopp J, Bartholdt L, Beier L, Liebert A 2007. Second to fourth digit length ratio 2D:4D. and adult sex hormone levels: New data and a meta-analytic review. *Psychoneuroendocrinology* 32: 313–321. doi:10.1016/j.psyneuen.2007.01.007.

Hönekopp J, Watson S 2010. Meta-analysis of digit ratio 2D:4D shows greater sex difference in the right hand. *Am J Hum Biol* 22: 619–630. doi:10.1002/ajhb.21054.

Hönekopp J, Watson S 2011. Meta-analysis of the relationship between digit-ratio 2D: 4D and aggression. *Personality and Individual Differences* 51: 381–386.

Hönekopp J, Voracek M, Manning JT 2006. 2nd to 4th digit ratio 2D:4D. and number of sex partners: Evidence for effects of prenatal testosterone in men. *Psychoneuroendocrinology* 31: 30–37. doi:10.1016/j.psyneuen.2005.05.009.

Jäckle, Annette, Peter Lynn, and Jon Burton. 2013. Going Online with a Face-to-Face Household Panel: Initial Results from an Experiment on the *Understanding Society* Innovation Panel. *Understanding Society Working Paper Series No. 2013 – 03* Institute for Social and Economic Research, University of Essex

James WH 2001. Finger-length ratios, sexual orientation and offspring sex ratios. *J Theor Biol* 212: 273–274. doi:10.1006/jtbi.2001.2376.

Kaminska, Olena and Peter Lynn 'Tailoring mode of data collection in longitudinal studies.' In *Understanding Society Innovation Panel Wave 5: Results from Methodological*

Experiments. UKHLS working paper 2013-06. (Ed. Jon Burton), ISER, Colchester, University of Essex.

Kennickell, Aurthur B. 1996. .Using Range Techniques with CAPI in the 1995 Survey of Consumer Finances.. *Paper presented at the Joint Statistical Meetings, August 1995: Chicago, IL.*

Koenker, Roger and G. Bassett. 1978. Regression Quantiles. *Econometrica* 46:33-50

Chang, L. and Krosnick, J. A. 2010. Comparing oral interviewing with self-administered computerized questionnaires: An experiment. *Public Opinion Quarterly*, 74, 154-167.

Lippa RA 2003. Are 2D: 4D finger-length ratios related to sexual orientation? Yes for men, no for women. *Journal of personality and social psychology* 85: 179.

Manning JT, Scutt D, Wilson J, Lewis-Jones DI 1998. The ratio of 2nd to 4th digit length: A predictor of sperm numbers and concentrations of testosterone, luteinizing hormone and oestrogen. *Hum Reprod* 13: 3000–3004. doi:10.1093/humrep/13.11.3000.

Moore, J.C., L.L. Stinson, and E.J. Welniak. 1999. .Income Reporting in Surveys: Cognitive Issues and Measurement Error.. in *Cognition and Survey Research*, edited by M. G. Sirken, D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur, and R. Tourangeau. New York: Wiley.

Nelson RJ 2011. *An introduction to behavioral endocrinology 4th edition.*. Sinauer. 670 p.

Palta, Mari, Ronald J. Prineas, Reuben Berman, and Peter Hannan. 1982. Comparison of self-reported and measured height and weight. *American Journal of Epidemiology* 115:223-230.

Porst, Rolf and Klaus Zeifang. 1987. A Description of the German General Social Survey Test-Retest Study and a Report on the Stabilities of the Sociodemographic Variables. *Sociological Methods and Research* 15:177-218.

Putz DA, Gaulin SJC, Sporter RJ, McBurney DH 2004. Sex hormones and finger length: What does 2D:4D indicate? *Evolution and Human Behavior* 25: 182–199. doi:10.1016/j.evolhumbehav.2004.03.005.

Rowland, Michael L. 1990. Self-Reported Weight and Height. *American Journal of Clinical Nutrition* 52:1125-1133.

Shanahan MJ, Hofer SM, Shanahan L 2003. Biological models of behavior and the life course. In: Mortimer JT, Shanahan MJ, editors. *Handbook of the life course*. New York: Kluwer Academic Publishers.

Spencer, Elizabeth A., Paul N. Appleby, Gwyneth Davey, and Timothy J. Key. 2002. Validity of self-reported height and weight in 4808 EPIC-Oxford participants. *Public Health Nutrition* 5:561-565.

Stewart, Alistair W., Rodney T. Jackson, Michael A. Ford, and Robert Beaglehole. 1987. Underestimation of relative weight by use of self-reported height and weight. *American Journal of Epidemiology* 125:122-126.

Sturgis, Patrick, Nick Allum, and Ian Brunton-Smith. 2009. Attitudes Over Time: The Psychology of Panel Conditioning. Pp. 113-126 in *Methodology of Longitudinal Surveys*, edited by P. Lynn. New York: John Wiley & Sons, Ltd.

Taylor CJ 2014. Physiological stress response to loss of social influence and threats to masculinity. *Social Science & Medicine* 103: 51–59. doi:10.1016/j.socscimed.2013.07.036.

Tourangeau, Roger, Ken Rasinski, J.B. Jobe, T.W. Smith, and W. Pratt. 1997. Source of Error in a Survey of Sexual Behaviour. *Journal of Official Statistics* 13:341-365.

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

Traugott, M.W. and J.P. Katosh. 1979. Response validity in surveys of voting behavior. *Public Opinion Quarterly* 43:359-377.

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

Waterton, Jennifer and Denise Lievesley. 1989. Evidence of conditioning effects in the British Social Attitudes Panel. Pp. 319-339 in *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh. New York: John Wiley & Sons.

Van der Vaart, W. 2009. Enhancing Long-Term Memory for Attitudes in Survey Research: Can a Cue-List Help? A Field Experiment on Aided Recall. *Journal of Official Statistics*, 25 363-378.

Voracek M, Pum U, Dressler SG 2010. Investigating digit ratio 2D:4D. in a highly male-dominated occupation: The case of firefighters. *Scandinavian Journal of Psychology* 51: 146–156. doi:10.1111/j.1467-9450.2009.00758.x.

Voracek M, Loibl LM 2009. Scientometric analysis and bibliography of digit ratio 2D:4D. research, 1998-2008. *Psychological Reports* 104: 922–956. doi:10.2466/pr0.104.3.922-956.

Voracek M, Pietschnig J, Nader IW, Stieger S 2011. Digit ratio 2D:4D. and sex-role orientation: Further evidence and meta-analysis. *Personality and Individual Differences* 51: 417–422. doi:10.1016/j.paid.2010.06.009.

Voracek M, Tran US, Dressler SG. 2010. Digit ratio 2D:4D. and sensation seeking: New data and meta-analysis. *Personality and Individual Differences* 48: 72–77. doi:10.1016/j.paid.2009.08.019.