Ask a silly question and get a silly answer?

An experimental analysis of the impact of survey design on measures and models of subjective wellbeing

Angus Holford

Stephen Pudney

Institute for Social and Economic Research
University of Essex

This version: December 20, 2013

Abstract

We analyse the results of experiments on aspects of the design of questionnaire and interview mode in the first four waves (2008-11) of the UK *Understanding Society* panel survey. The randomised experiments relate to job, health, income, leisure and overall life-satisfaction questions and vary the labeling of response scales, the mode of interviewing and the location of questions within the interview. We find significant evidence of an influence of interview mode and (to a lesser extent) question design on the distribution of reported satisfaction and self-assessed health, particularly for women. Results from the sort of conditional modeling used to address real research questions are also affected to some extent, but appear less vulnerable to design influences than simple summary statistics.

Keywords: Survey design, wellbeing, satisfaction, response bias, Understanding Society

JEL classifications: C23, C25, C81, J28

Contact: Steve Pudney, ISER, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK; tel. +44(0)1206-873789; email spudney@essex.ac.uk

This work was supported by the UK Longitudinal Studies Centre (award no. RES-586-47-0002) and the European Research Council (project no. 269874 [DEVHEALTH]). We are grateful to Jon Burton, Anette Jäckle and Noah Uhrig for their help with the IP data and to participants at the LSE seminar in wellbeing measurement March 2012 for helpful discussion.

1 Introduction

Subjective assessments of wellbeing play an increasingly important role in applied economics. They underpin some approaches to economic evaluation, for example in assessment of health technologies and interventions (Ferrer-i-Carbonell and van Praag 2002); they have been proposed for legal assessment of compensation claims in the courts (Oswald and Powdthavee 2008), and they have been used as indicators of developmental outcomes and some aspects of the underlying non-cognitive skills emphasised by Heckman et al (2006). In political circles, wellbeing is also in the air. In France, President Nicolas Sarkozy's *Commission sur la Mesure de la Performance Économique et du Progrès Social* of 2008-9 (Stiglitz et al 2009) proposed measures of economic performance with wider scope than traditional indicators like GDP. In November 2010, the British Prime Minister announced plans for the Office for National Statistics to develop and publish official measures of wellbeing, observing that "prosperity alone can't deliver a better life" (Cameron 2010). Other national governments and international organisations including the OECD, Eurostat and the UN have made similar moves to extend the range of welfare indicators they produce. Much of the discussion about wellbeing measurement in the economics literature has focused on conceptual issues relating to distinctions between satisfaction, happiness, positive and negative affect, etc, distinctions between different domains of wellbeing (Stiglitz et al 2010), and validating international comparisons of the level and determinants of wellbeing (Kapteyn et al 2013). More practical questions about survey design for subjective questions have long been discussed in the survey methods literature, but have not featured much in the economic debate over wellbeing measurement.

Everyone knows that the way you ask a question may have a big influence on the answer that you get, and subjective questions on health and wellbeing are no exception to this. Perhaps more important than the question itself is the form in which you require the answer – and evidence of difficulty in interpreting response scales has been found in qualitative work by Ralph et al (2011). Although there has been interest in reliability issues for subjective wellbeing data (see, for example, Kristensen and Westergaard-Nielsen 2007 and Krueger and Schkade 2008), the economics literature on happiness and wellbeing has not devoted much attention to the influence of survey design and context and its possible implications for data analysis. However, Conti and Pudney (2011) analysed quasi-experimental evidence arising from variations over time in the design of job satisfaction questions in the British Household Panel Survey (BHPS), finding evidence of a substantial influence of the design of questions and response scales, the mode of interview and the interview context on the distribution of reported satisfaction. One of the most striking aspects of the BHPS evidence is the significantly different impact that survey design features have on the response behaviour of men and women and the consequent large distortions that may be induced in research findings on gender differences in the determinants of job satisfaction. The lack of robustness of gender differences was also found by Studer (2012), who analysed a randomised experiment comparing continuous and discrete rating scales in a Dutch web-based panel survey.

Although choice of question design and interview mode have been examined by survey methodologists, their conclusions are typically limited to simple indicators of data quality and impacts on summary statistics like means and sample proportions. But, in practice, measures of wellbeing are used in much more sophisticated ways, for conditional modeling and comparison over time and across groups within society. The political drivers of the move to measurement are quite clear about this: "If anyone was trying to reduce the whole spectrum of human happiness into one snapshot statistic, I would be the first to roll my eyes [...]. But that's not what this is all about" (Cameron 2010).

In the economics literature, claims for the validity of subjective wellbeing indicators are often based on a loose prediction argument: that these variables are correlated in the way one would expect with observable variables and with subsequent behavioural outcomes. The predictive criterion is an important one, but not very powerful. It is easy to construct examples where an indicator has measurement error serious enough to cause catastrophic biases for the sort of analysis economists are interested in, but is sufficiently highly correlated with the 'true' variable to satisfy the requirement for predictive correlation. Our aim in this paper is to add to the evidence on measurement reliability, by using a set of randomised experiments to investigate the impact of various dimensions of survey design in the context of a major UK survey that is used for wellbeing analysis: the UK Household Longitudinal Survey (UKHLS), also known as *Understanding Society*, which is the successor to the BHPS. This paper extends Conti and Pudney's (2011) BHPS analysis of job satisfaction by considering also self-reported health and several other domains of life satisfaction, and by using experimental control and a wider range of variations in survey design. In section 5, we also look at related experiments focused on self-reported health.

2 The UKHLS Innovation Panel

The UKHLS is a very large-scale, household panel survey, which has absorbed the long-established BHPS sample. Its design differs in many ways from that of BHPS and one of its innovative features is a sub-panel, known as the *Innovation Panel* (IP), reserved exclusively for experimental work. The IP experiments constitute one of the very few examples of experimentation in survey design sustained over a substantial number of waves of a panel study; there is an annual competition open to all researchers to propose new experiments.[1] The IP sample of 1500 households was drawn in Spring 2008 and has been re-interviewed annually since then. There is a relatively high attrition rate (McFall et al 2013), so a refreshment sample of 500 households was added at wave 4 in Spring 2011. The core content of the IP interview is similar to the UKHLS main survey, but there are considerable variations in content from wave to wave. Each wave carries several experiments: during waves 1-4, an annual average of three procedural experiments (interview mode, incentives, etc) and five measurement experiments (questionnaire format, wording, etc). As a consequence, the

[1] The German Socio-Economic Panel (GSOEP) has also now developed an IP sub-panel of the GSOEP.

observed outcomes are affected by multiple interventions and the complicated nature of the sample that results is a disadvantage of the IP concept. However, each experiment is randomized, so cross-contamination between experiments can be expected to contribute variance rather than bias. The major advantage of the IP is that experiments take place within the context of a large, continuing panel survey, so the IP is believed to be superior in terms of external validity to the small special-purpose experiments which are typical of much of the survey methods literature.

2.1 Experimental design: treatment groups

Fieldwork for waves 1-4 of the IP were conducted in April-June in Spring of each of the four years 2008-2011, and included experiments to investigate the impact of question design, interview mode and positioning of questions within the interview, using random assignment of households to treatment groups. All individuals within the household received the same experimental treatment. In this study, we focus on questions covering five satisfaction domains: a 4-item module covering satisfaction with health, income, leisure and life overall, and a separate job satisfaction question. Each measure involves a 7-point response scale, as used in the BHPS from 1991 to 2008. This excludes all wave 1 health/income/leisure/life satisfaction data and a random half of the wave 1 job satisfaction responses, which used an 11-point scale (see Burton et al (2008) for a comparison of IP responses using 7-point and 11-point scales).

Wave 1 A random half of the sample received a job satisfaction question with a 7-point scale (question wordings are given in section 2.2). The question was delivered orally by an interviewer following a computer-generated script without use of a showcard. Verbal equivalents ("completely dissatisfied" and "completely satisfied") were given only for the two extreme points on the scale.

Wave 2 The wave 2 design was a composite of two separate randomisations. First, households were assigned in the ratio 2:1 to groups interviewed by Computer-Assisted Telephone Interviewing (CATI) or face-to-face (F2F) during an interviewer visit to the home. During F2F interviews, most questions were delivered by Computer-Assisted Personal Interviewing (CAPI), but Computer-Assisted Self-Interviewing (CASI) was used for the satisfaction module in a randomly-assigned subgroup. There were also independent assignments to treatment groups formed by varying question design and position of the question within the interview. As part of the design, for assignments that would have resulted in a requirement to administer by CATI a question that was in fact infeasible by telephone (because it required a showcard or reading of a long list of allowable responses), the closest feasible approximation to the allocated treatment was substituted. In some cases telephone contact was unsuccessful, in which case some or all members of the household were instead interviewed F2F, if that proved possible. There were no variations in question position within the interview for job satisfaction, so there were 8 treatment groups at wave 2 for job satisfaction, rather than the 14 for other domains.

Wave 3 Wave 3 was conducted entirely F2F, so there were no CATI groups. Apart from that, the set of experimental treatments was identical to those used at wave 2, but a group rotation was used to generate temporal variation in treatments. There was also a further – unintended – experiment at wave 3, since an error in the CAPI script led to some questions being repeated in a different format within the same interview. The impact of that inadvertent repetition is examined in section 3.1.

Wave 4 For the 4-item satisfaction module, the wave 4 experiment was a simple comparison of two private modes: a paper self-completion (PSC) questionnaire and a CASI version. The job satisfaction question was administered to all employed respondents in standard CAPI mode.

The longitudinal pattern of treatments affecting the IP satisfaction modules is detailed in Tables 1 and 2, together with potential sample numbers on an intention-to-treat (ITT) and an actual response basis. To avoid difficult selection issues, all the results presented here are on an ITT basis, although there is very little difference in the findings when the analysis is repeated using actual treatments, owing to the high compliance rate in most groups. Most of the other experiments carried by the IP in waves 1-4 were irrelevant to our objectives, but one of them – an experiment with the wording of a subjective health question – is allowed for as a further 'treatment' when we examine health measurement in section 5.

Table 1 Experimental treatment groups: satisfaction variables with 7-point response scale

	Treatment group	**Question placement**	**Sample size**							
			Wave 1		Wave 2		Wave 3		Wave 4	
			ITT[1]	Actual[2]	ITT[1]	Actual[2]	ITT[1]	Actual[2]	ITT[1]	Actual[2]
	Satisfaction with Health, Income, Leisure and Life overall									
1	Visit: CASI full labels	Late	-	-	126	167	135	123	1,208	1,060
2	Visit: CASI polar labels	Late	-	-	110	136	306	286	-	-
3	Visit: showcard full labels	Late	-	-	47	55	291	270	-	-
4	Visit: oral 2-stage	Late	-	-	59	80	286	260	-	-
5	Visit: showcard polar labels	Late	-	-	48	69	424	391	-	-
6	Visit: oral polar labels	Late	-	-	63	76	314	287	-	-
7	Phone: 2-stage	Late	-	-	404	297	-	-	-	-
8	Phone: polar labels	Late	-	-	402	308	-	-	-	-
9	Visit: showcard full labels	Early	-	-	58	65	-	-	-	-
10	Visit: oral 2-stage	Early	-	-	56	69	-	-	-	-
11	Visit: showcard polar labels	Early	-	-	52	56	-	-	-	-
12	Visit: oral polar labels	Early	-	-	59	74	-	-	-	-
13	Phone: 2-stage	Early	-	-	188	140	-	-	-	-
14	Phone: polar labels	Early	-	-	198	161	-	-	-	-
15	Visit: PSC full labels	Separate	-	-	-	-	-	-	1,177	855
	Job satisfaction									
1	Visit: CASI full labels	Mid	-	-	76	101	159	140	-	-
2	Visit: CASI polar labels	Mid	-	-	61	78	170	161	-	-
3	Visit: showcard full labels	Mid	-	-	60	61	169	155	1,380	1,247
4	Visit: oral 2-stage	Mid	-	-	64	87	159	149	-	-
5	Visit: showcard polar labels	Mid	-	-	51	67	155	141	-	-
6	Visit: oral polar	Mid	731	672	64	79	158	140	-	-
7	Phone: 2-stage	Mid	-	-	331	246	-	-	-	-
8	Phone: polar labels	Mid	-	-	322	256	-	-	-	-

[1] Intention-to-treat: assigned treatment [2] Treatment received by responders

2.2 Experimental design: question wording and response scales

Questions were asked sequentially for three aspects of life satisfaction, using (for all except groups 4, 7, 10 and 13) the following question stem:

> *How dissatisfied or satisfied are you with the following aspects of your situation: (a) your health; (b) the income of your household; (c) the amount of leisure time you have.*

These three domain-specific questions were followed by an overall assessment:

> *Using the same scale, how dissatisfied or satisfied are you with your life overall?*

For groups 3 and 9, a fully-labeled showcard specified response options in a vertical list ordered from top to bottom as: *7 Completely satisfied; 6 Mostly satisfied; 5 Somewhat satisfied; 4 Neither satisfied nor dissatisfied; 3 Somewhat dissatisfied; 2 Mostly dissatisfied; 1 Completely dissatisfied.* For groups 1 and 2, the questions were administered by the more private Computer-Assisted Self-Interviewing (CASI) method, and the seven alternatives were displayed horizontally across the screen of a laptop computer for selection directly by the respondent. Polar-point labeled variants of the question (groups 5, 6, 8, 11, 12 and 14) omitted the textual labels from options 2 to 6. If the polar-labeled response scale was communicated orally (groups 6, 8, 12 and 14), explanations of the two extreme points were read out by the interviewer.

Groups 4, 7, 10 and 13 received a question with a fully-labeled response scale, designed to be deliverable by telephone, when use of a showcard or reading of a full list of responses would have been impractical. The question has a two-stage structure:

> (i) *How dissatisfied or satisfied are you with [...]? Would you say you are...*
> *(1 Dissatisfied; 2 Neither Dissatisfied nor Satisfied; 3 Satisfied)*

(ii) [If dissatisfied or satisfied...] *And are you Somewhat, Mostly or Completely [Satisfied / Dissatisfied] with [...]? (1 Somewhat; 2 Mostly; 3 Completely)*

At wave 2, questions on satisfaction with health, family income, leisure and life overall were either asked early (about 25% of the way through the interview, following a block of questions on transport mode choices) or late (about 95% of the way through the interview, following questions on political affiliation and values). At waves 3 and 4, these questions were always asked late, except for group 15 at wave 4, where the questions were contained in a paper self-completion questionnaire completed during the interviewer's visit.

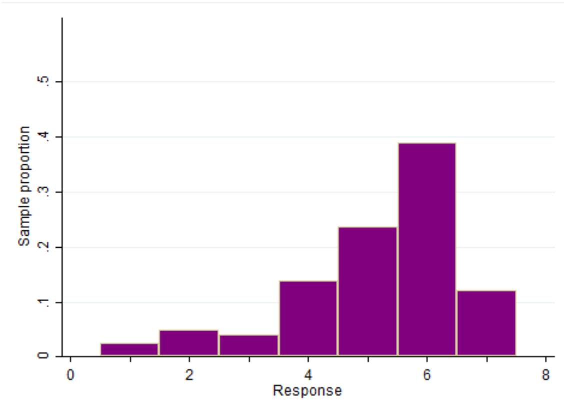
People in employment or self-employment were also asked about their job satisfaction: shortly after mid-interview, following a section dealing with employment or self-employment details, including occupation, hours and earnings. The question stem is:

All things considered, which number best describes how dissatisfied or satisfied you are with your job overall?

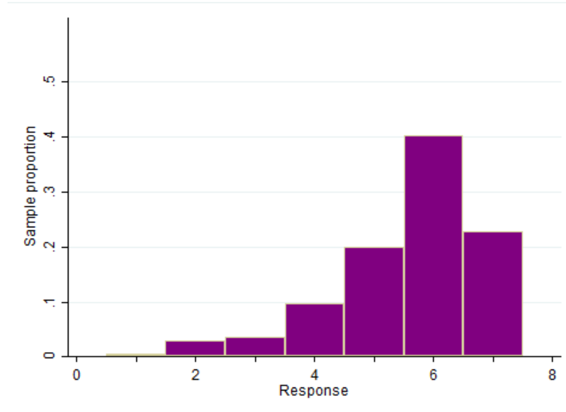
The same 1-7 response scale and labeling options were used as for the single-stage life satisfaction questions, and groups 4 and 7 received a 2-stage variant.

3 The impact on data distributions

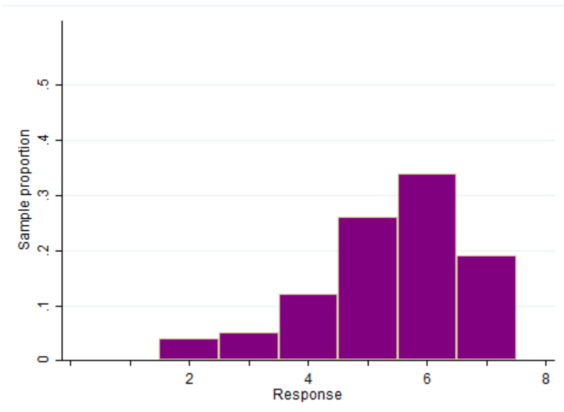
We are mainly interested in the effect of survey design on the answers to substantial empirical research questions; but we first look for evidence that the distribution of responses to questions on subjective wellbeing are influenced by aspects of design. Appendix Table A1 gives the mean responses for each treatment group, separately for men and women, together with wave-specific tests comparing the distribution of responses from each treatment group with the distribution in the rest of the sample. There is some evidence of an impact of the set of experimental variations, but the small group sizes mean that tests for individual treatment groups have low power. Nevertheless, at waves 2 and 3, Table A1 suggests a definite pattern for CASI compared to other more public interview modes: looking across all five satisfaction domains and both forms of CASI, 18 of the 20 mean scores are below the overall average for women and 14 of 20 are below average for men. Figure 1 compares the distributions for CASI responses to the life satissfaction question with other one-stage F2F designs at wave 2. The distributions are dominated by a mode at $Y = 6$, which is a general feature of categorical responses to satisfaction questions, possibly reflecting an aversion to extremes, as suggested by Studer (2012). The comparison of CASI with other designs suggests a shift of mass from $Y = 6$ and 7 to $Y \leq 4$: overall, CASI increases the sample proportion of $Y \leq 4$ from 16% to 23% and reduces the sample proportion of $Y \geq 6$ from 61% to 52%. It is not a simple matter to interpret these mode effects, since they involve differences in several dimensions, including the format of visual display of the response scale (Jenkins and Dillman 1997), the degree of respondent privacy and presence of an outsider (the interviewer). Privacy and the social desirability of alternative responses are especially important for sensitive issues (Hochstim 1967, De Leeuw 1992, Aquilino 1997) and a further important factor may be a desire by some individuals to maintain a bargaining position within the family, rendering some satisfaction questions sensitive in oral interviews where other family members may be within earshot (Conti and Pudney 2011).



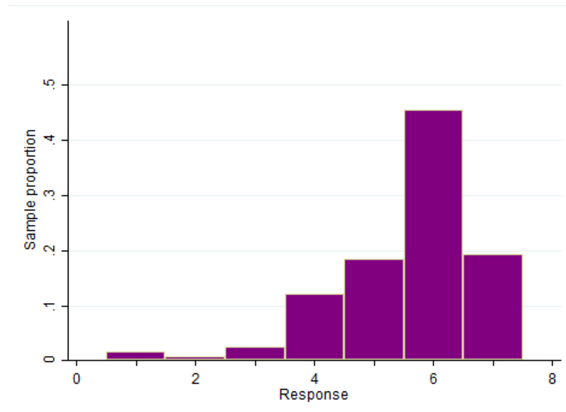
(a) CASI: women ($n = 126$)



(b) Other F2F 1-stage designs: women ($n = 295$)



(c) CASI: men ($n = 110$)



(d) Other F2F 1-stage designs: men ($n = 230$)

Figure 1 Wave 2 sample distributions for life satisfaction: CASI vs. 1-stage F2F

Table 2 gives results from overall tests of joint significance for the whole set of experimental variations, by survey wave and domain of satisfaction. In view of earlier findings on gender differences, we carry out these tests separately for men and women. Let Y_d be the satisfaction score for the dth domain. We use Monte Carlo permutation versions of the Kruskal-Wallis H-statistic and ANOVA F-statistic for Y_d (see Good 2006 for a review of permutation tests and Heckman et al 2010 for an application to experimental evaluation).

Table 2 Permutation test P−values for joint hypothesis of no treatment effects

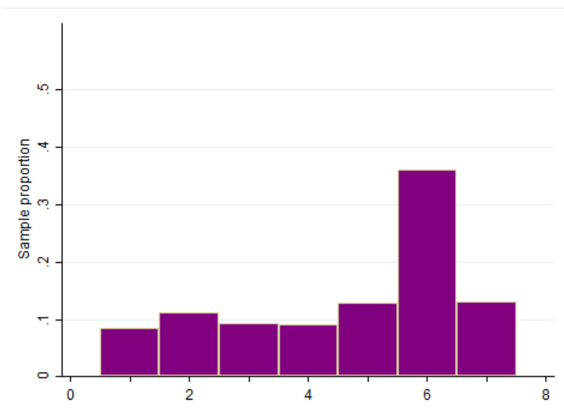
	Women			Men		
Satisfaction domain	wave 2	wave 3	wave 4	wave 2	wave 3	wave 4
Health	0.053	0.250	0.066	0.227	0.155	0.003
	0.000	*0.089*	*0.329*	*0.047*	*0.027*	*0.008*
Income	0.240	0.350	0.094	0.219	0.929	0.223
	0.037	*0.182*	*0.310*	*0.029*	*0.889*	*0.884*
Leisure	0.176	0.359	0.107	0.150	0.196	0.487
	0.024	*0.269*	*0.207*	*0.022*	*0.105*	*0.716*
Life overall	0.000	0.604	0.393	0.059	0.299	0.261
	0.000	*0.139*	*0.291*	*0.063*	*0.190*	*0.677*
Job	0.007	0.359	.	0.058	0.245	.
	0.000	*0.052*	.	*0.063*	*0.373*	.

Roman: based on Kruskal-Wallis H statistic for Y_d. *Italic*: ANOVA F−statistic. 10,000 replications

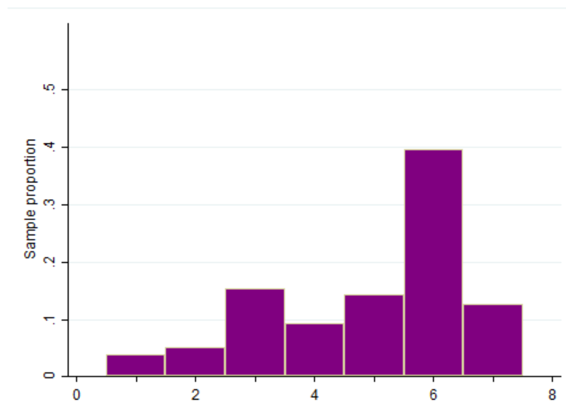
The test results of Table 2 show that design effects are frequently significant, although the pattern of effects is unexpected. Experimental variation at wave 2 produces large impacts on the response distributions: for women, they are significant in all five domains at the 5% level using an ANOVA permutation F test, and in two of the five domains using a permutation H-statistic. For men, we find a significant ANOVA test in the majority of domains and some evidence at the 10% level for the H-statistic. But at wave 3, where group sizes are larger and we would expect better power, there is very little evidence of an effect: for women, only in two of the domains at the 10% level and for men a single 5% rejection (satisfaction with health) in the ANOVA test, with no rejections at all by the H-statistic. A possible interpretation of the lack of effect at wave 3 is linked to the rotation of treatment groups betweeen waves 2 and 3. Since almost every wave 3 respondent had responded via a different mode or question design a year earlier, the recollection of that response may have nullified the effect of treatment at wave 3 – which would be consistent with Pudney's (2008, 2011) findings of dynamic contamination of responses to a different subjective wellbeing question in the BHPS. If that explanation is accepted, then it casts doubt on the validity of observed measures of change in wellbeing in panel data.

The experimental treatment groups differ in a number of dimensions, and tests of the impact of specific aspects of design (rather than combinations of aspects) are more informative. Appendix Table A2 gives the results of permutation tests for contrasts between three aspects of question design: late/early placement, full/polar labeling of response scales, and 2-stage/1-stage response scales; and three interview modes: CASI/PSC, CASI/F2F and CATI/F2F. These tests are implemented in a context-specific way, separately for different combinations of the other design aspects. Again, this involves small sample sizes and relatively low power. However, at wave 2, there is some evidence of an impact of the 2-stage rather than 1-stage question design (mainly for female respondents) and, more tentatively, of full rather than polar-point labeling of the response scale. There is also some weak wave 2 evidence of an effect of the F2F rather CASI mode. Although wave 3 shows little sign of any design effect, there is evidence of an impact of CASI against the F2F mode for the health and leisure domains (men) and income (women). It could be argued that this represents a privacy effect, with men reluctant to express public dissatisfaction with their health or leisure and women reluctant to voice concerns about income. In all of these cases, CASI delivers a significantly lower mean satisfaction score.

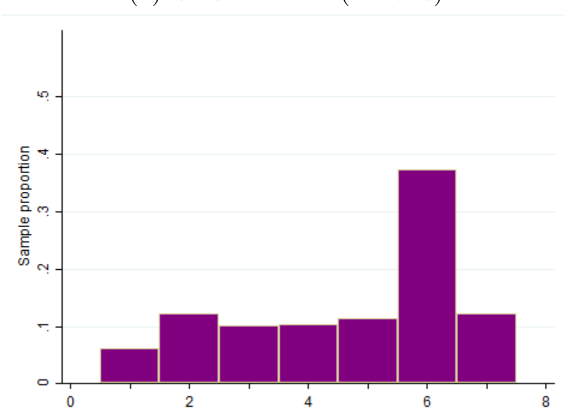
At wave 4, there is a significant effect for CASI rather than PSC, especially for satisfaction with health among male respondents, for whom CASI produces a much smaller mean response (4.70) than PSC (5.12). Figure 2 shows the wave 4 response distributions for satisfaction with health, by gender and interview mode. Compared with PSC, CASI has the effect of transferring probability mass to categories $Y = 1$ and 2, from $Y = 6$ in particular. This reduces the mean score, but also changes the mass of the lower tail, which has implications for the common applied practice of using binary indicators of low satisfaction. The impact on the response distribution is surprising because CASI and PSC are both private modes designed to do essentially the same thing: shield the respondent from social pressures during interview. Assuming they both achieve that aim, the remaining difference between them must presumably relate to the way in which the response scale is conveyed on the computer screen or paper questionnaire and then interpreted by the respondent. However, both use the same fully-labeled response scale. In CASI they are displayed vertically from 1 = completely dissatisfied at the top of the screen to 7 = completely dissatisfied at the bottom, whereas the paper questionnaire displays them hozontally from 1 at the left to 7 at the right. The significant differences we find are consistent with the warning from Christian et al (2009) that the visual design of response scales can have a significant influence on responses. It is likely to become a particular issue in future multi-mode surveys which have difficulty in avoiding endogenous selection from the set of interview modes, each of which has a distinct 'look'.



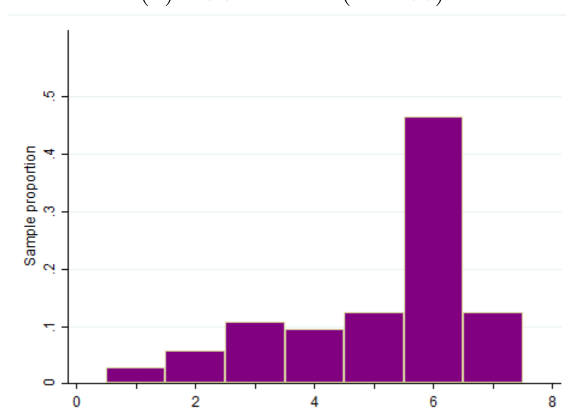
(a) CASI: women (n = 578)



(b) PSC: women (n = 496)



(c) CASI: men (n = 578)



(d) PSC: men (n = 496)

Figure 2 Wave 4 sample distributions for satisfaction with health: CASI vs. PSC

To clarify the picture, Table 3 reports the results of extending simple ANOVA comparisons to the health, income, leisure and life satisfaction domains, using a seemingly unrelated regressions approach allowing unrestricted correlation between the four satisfaction scores. The analysis is restricted to wave 2 (the analogous estimates for wave 3 show little impact) and focuses on two interview mode contrasts (CASI v. F2F and CATI v. F2F) and two question design constrasts (polar-point v. full labeling of the response scale and 2-stage v. 1-stage question design). The analysis is applied within subsamples which have approximately the same composition in terms of all other experimental aspects for the two groups being compared, so that there should be negligible compositional bias in the comparisons reported. For comparison, we also include analogous single-equation results for the smaller group of employed respondents who are asked a separate job satisfaction question. For each panel of Table 3, the first four rows show the mean effects on domain-specific satisfaction scores; the fifth row gives a joint P-value for the joint hypothesis that all four mean effects are zero. The clearest evidence from these joint tests is for CASI rather than F2F interviewing and for 2-stage rather than 1-stage question design, but both results apply only to female respondents. Evidence on job satisfaction shows the same pattern.

Table 3 IP wave 2: Impacts on mean responses of specific design aspects

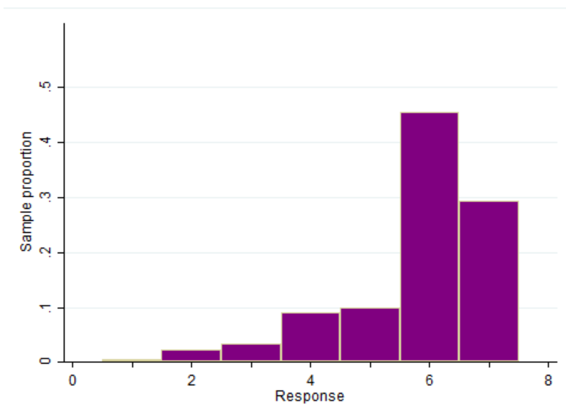
Satisfaction domain	interview mode		Response scale design	
	CASI[1]	CATI[2]	Polar-point[3]	2-Stage[4]
Women				
Health	-0.225	0.109	-0.044	0.132
	(0.222)	(0.170)	(0.222)	(0.122)
Income	0.069	0.008	−0.438**	0.152
	(0.221)	(0.174)	(0.219)	(0.125)
Leisure	-0.340	0.173	-0.328	0.374***
	(0.223)	(0.185)	(0.223)	(0.133)
Overall	−0.627***	0.123	-0.253	0.255***
	(0.164)	(0.135)	(0.168)	(0.097)
Joint P-value[5]	0.0003	0.8020	0.2386	0.0244
n	227	727	227	727
Job[6]	−0.654***	0.334	-0.385	0.319**
	(0.267)	(0.210)	(0.272)	(0.149)
Men				
Health	0.008	0.112	−0.401*	-0.036
	(0.219)	(0.176)	(0.216)	(0.126)
Income	-0.158	0.174	0.147	0.031
	(0.237)	(0.188)	(0.236)	(0.134)
Leisure	-0.191	0.047	-0.093	0.031
	(0.268)	(0.205)	(0.267)	(0.147)
Overall	−0.334*	-0.088	-0.172	0.124
	(0.187)	(0.153)	(0.188)	(0.109)
Joint P-value[5]	0.4009	0.5873	0.0696	0.7078
n	177	603	177	603
Job[6]	0.211	0.401*	-0.005	0.261*
	(0.318)	(0.224)	(0.313)	(0.157)

Standard errors in parentheses. Significance: *** = 1%; ** = 5%; * = 10%.
[1] Comparison with F2F interview + showcard: based on treatment groups 1-3, 5, 9, 11.
[2] Comparison with F2F oral (no showcard): based on treatment groups 4, 6-8, 10, 12-14.
[3] Comparison with fully-labeled scale: based on treatment groups 1-3, 5, 9, 11.
[4] Comparison with 1-stage question design: based on treatment groups 4, 6-8, 10, 12-14.
[5] SURE generalisation of the ANOVA test allowing for responses correlated across domains.
[6] Single-equation ANOVA test for subset of employed/self-employed individuals.

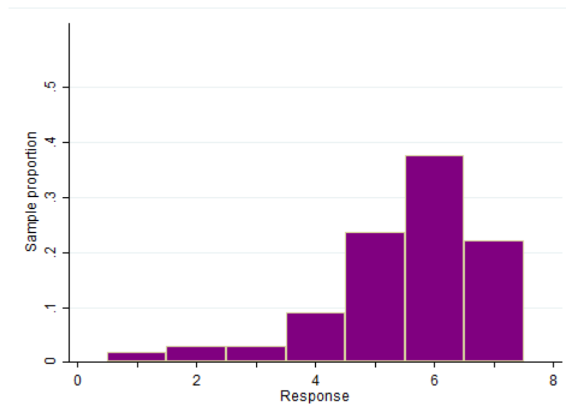
There has been some debate about the use of two-stage branching (or unfolding) question structures, some authors finding better reliability for them (Krosnick and Berent 1993),[2] while others found that some respondents have difficulty interpreting the question appropriately without access to the full range of allowed responses (Hunter 2005, p.10-11). Comparing

[2]Note that differences in question structure were confounded with labeling differences in the Krosnick-Berent study of test-retest reliability. We would also argue that test-retest reliability should be seen as a measure of consistency over time rather than 'reliability'.

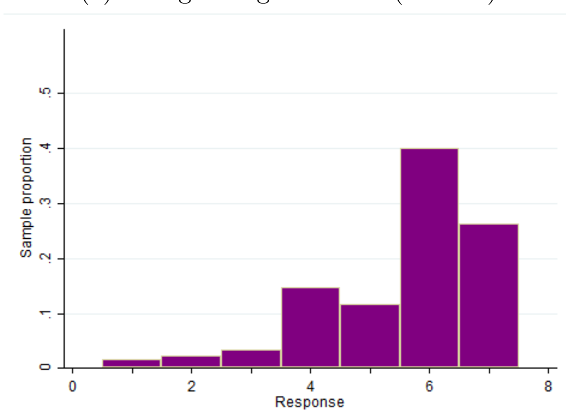
the 2-stage design with 1-stage alternatives in Table A1, we find higher mean scores for the 2-stage design in 16 out of 18 cases for women and 12 out of 18 for men. Again, Tables A2 and 3 show that these differences are statistically significant for women (leisure and life overall) but not men. Figure 3 shows the empirical response distributions and suggest that the main effect of the 2-stage design is to move responses from the $Y = 5$ category to $Y = 6, 7$, thus raising the mean score. There is little evidence of any difference between the 1-stage and 2-stage designs at wave 3.



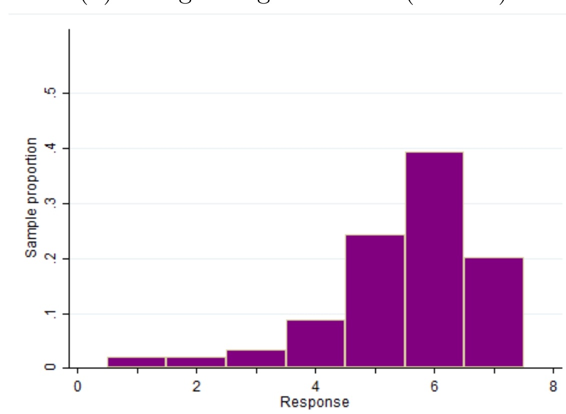
(a) 2-stage design: women ($n = 379$)



(b) 1-stage designs: women ($n = 641$)



(c) 2-stage design: men ($n = 328$)



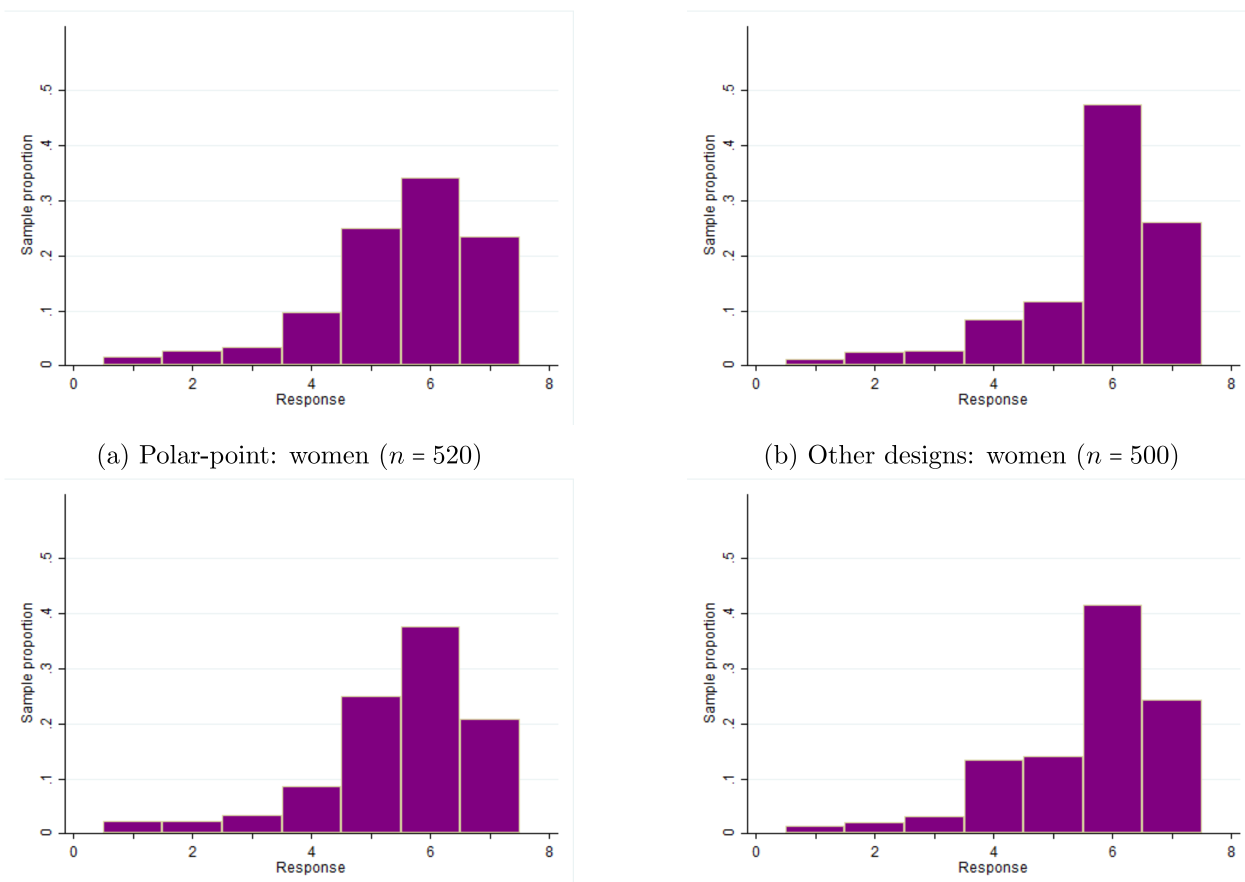
(d) 1-stage designs: men ($n = 522$)

Figure 3 Wave 2 sample distributions for life satisfaction: 2-stage vs. 1-stage question designs

We find no evidence that early or late positioning of questions within the questionnaire causes any significant shifts in the response distribution. This is in contrast to the large questionnaire context effects found in some other survey applications (Schuman and Presser

1981, Tourangeau 1999) and the evidence of respondent fatigue which may affect responses late in the interview (Herzog and Bachman 1981, Helgeson and Ursic 1994). Note that we do not investigate the ordering of individual questions within the satisfaction module – something that has been found to influence respondents' interpretation of satisfaction questions (Schwarz et al 1991, Tourangeau et al 1991).

Unlike Weng (2004) and Conti and Pudney (2011), there is only weak evidence of an impact of polar point rather than full labeling of the response scale (Tables A2 and 3). Its impact on mean scores is negative in most cases (Table 3), resulting from a shift from responses at $Y = 6$ to $Y = 5$ (Figure 4). This effect is surprising, given our expectation that exclusive labeling of extreme points would attract responses to those extremes.



(a) Polar-point: women ($n = 520$)

(b) Other designs: women ($n = 500$)

(c) Polar-point: men ($n = 412$)

(d) Other designs: men ($n = 438$)

Figure 4 Wave 2 sample distributions for life satisfaction: Polar-point vs. other question designs

3.1 Repeated measures within wave 3

Some respondents at wave 3 received the health, income, leisure and life satisfaction questions in two different forms within the same interview. This was an error in programming the CAPI script, rather than a deliberate experiment, but it offers a direct opportunity to assess the effects of different treatments on the same set of people. This allows us to compare responses to different designs more efficiently than through random assignment to single treatments. However, if the fact of repetition changes behaviour directly, or if there are significant question order or respondent fatigue effects, the results will be confounded to some extent.

Four groups received repeated questions. Group I received the 2-stage question, delivered orally early in the interview and then the single-stage question with fully-labeled showcard about 20 minutes later on average. Group II received the single-stage question orally, with verbal descriptions of the two extreme points, then later the same question using a polar-labeled showcard. Groups III and IV had the same treatments as I and II in reverse order. The top panel of Table 4 gives correlations between early and late scores and estimates of the mean differences. The test-retest correlations are in the range 0.57-0.86, which is rather higher than the range of correlations for life satisfaction quoted by Andrews and Whithey (1976), Kammann and Flett (1983) and Krueger and Schkade (2008), who used longer retest intervals but unchanged question design. If we make classical measurement error assumptions, the correlation between early and late measures gives the usual measure of test-retest reliability as the share of measurement error in total variance: implying a range of values of 0.16-0.75 for the noise/signal ratio (of the measurement error variance to the variance of the 'true' variable).

We investigate differences in the early and late mean scores and in the proportion of high scores ($Y \geq 6$ or $Y = 7$). For respondent i, the satisfaction score at time $t = 0$ (early) or 1 (late) is Y_{igt}, where g = group I, II, III or IV. At time t, members of groups I and II receive treatment sequences a, b and b, a respectively, while members of groups III and IV receive sequences c, d and d, c, where a, b, c, d denote the oral 2-stage question, fully-labeled showcard, oral polar-labeled question and polar-labeled showcard respectively. Assume additive effects:

$$Y_{igt} = \mu_0 + (\mu_a - \mu_b)\xi^a_{igt} + (\mu_c - \mu_d)\xi^c_{igt} + \mu_R(1-t) + \varepsilon_{igt} \tag{1}$$

where ξ^a_{igt}, ξ^c_{igt} are indicators of receiving treatments a and c respectively, and the ε^g_{it} are mutually independent zero mean measurement errors. The coefficient $(\mu_a - \mu_b)$ is the effect of using a 2-stage question structure rather than a showcard, $(\mu_c - \mu_d)$ is the effect of delivering the polar-labeled response question orally rather than by showcard, and μ_R is the effect of repetition. We estimate the coefficients by least squares random effects regression; the results are presented in the last panel of Table 4. We see significant effects for health satisfaction only, where use of the oral 2-stage question raises reported satisfaction relative to fully-labeled showcards, and question repetition has a positive effect of a similar magnitude.

Table 4 Repeated measures in IP wave 3

Treatment sequence	**Satisfaction domain** Health	Income	Leisure	Life
Test-retest Pearson correlation coefficients				
Oral 2-stage question → Fully labeled showcard[1]	0.665	0.707	0.672	0.571
Fully labeled showcard → Oral 2-stage question[2]	0.770	0.737	0.745	0.591
Polar labeled oral → Polar labeled showcard[2]	0.743	0.786	0.686	0.723
Polar labeled showcard → Polar labeled oral[3]	0.708	0.860	0.817	0.749
Mean scores: random effects regression				
Oral 2-stage question v. Fully labeled showcard: $(\mu_a - \mu_b)$	0.147*	-0.003	0.118	0.031
	(0.076)	(0.077)	(-0.085)	(0.064)
Polar labeled oral v. Polar labeled showcard: $(\mu_c - \mu_d)$	0.082	0.056	-0.120	-0.049
	(0.103)	(0.096)	(0.108)	(0.083)
Repetition effect: μ_R	0.143*	0.056	-0.006	0.012
	(0.077)	(0.078)	(0.087)	(0.064)
Sample size n =	512	503	511	512

[1] $n = 124$; [2] $n = 117$; [3] $n = 153$. Test statistics based on robust standard errors.

4 Survey design and satisfaction models

The demand for data is a derived demand – we are interested in data only because of the research results that can be produced from them. Much of the survey methods literature ignores this fundamental point and restricts consideration of the impact of design features to the statistical reliability of relatively simple summary measures computed from the data. Instead, most applied researchers are interested in the statistical relationships between variables, using models which represent complex conditional distributions in the data. In the research literature on wellbeing, this type of modeling takes the form of relationships between satisfaction as a dependent variable and a set of covariates describing the individual's characteristics and circumstances in some detail (see Van Praag and Ferrer-i-Carbonell 2004, and Clark et al 2008 for surveys). Typical analysis methods include fixed-effects regression and random-effects ordered probit. We apply these modeling approaches and investigate the impact of experimental variations in survey design on the estimates.

It is no simple matter to assess the impact of a set of experimental design variations on these complex analyses. With 15 treatment groups and models involving 20 or more coefficients, there are at least 3,000 experimental effects to be estimated in the most general approach. We resolve this 'curse of dimensionality' by focusing on the answers to specific research questions rather than model parameters. In this section, we consider two issues: in section 4.1, the possible gender difference in pecuniary influences on wellbeing; and, in section 4.2, the magnitude of the compensating income variation which would be required to offset the wellbeing effects of a persistent health condition. In both cases, we investigate the effect of using F2F interviewing rather than other, more private modes.

Two single-equation model specifications are used, both based on the following latent regression:

$$Y_{it}^{*} = \boldsymbol{x}_{it}\boldsymbol{\beta} + \boldsymbol{x}_{it}^{+}\zeta_{it}\boldsymbol{\gamma} + u_i + \varepsilon_{it} \tag{2}$$

where Y_{it}^{*} is the (latent) satisfaction score, $\boldsymbol{x}_i$ is the full vector of covariates, $\boldsymbol{x}_i^{+}$ is the subset of covariates of particular interest for a particular research question and ζ_i is a dummy indicating cases featuring a specific design aspect of interest. u_i and ε_{it} are unobservables. Let $\boldsymbol{T}_{it}$ be a vector of indicators describing the design treatment experienced by individual

i at time t; the observed score Y_{it} is then related to Y^*_{it} and $\boldsymbol{T}_{it}$ in alternative ways by the two models:

(i) *Fixed-effects (FE) regression*: $Y_{it} = Y^*_{it} + \boldsymbol{T}_{it}\boldsymbol{\alpha}$ and u_i is eliminated by removing within-group means.

(ii) *Generalised random-effects ordered (GREO) probit*: $Y_{it} = r$ iff $A^{r+1}_{it} \geq Y^*_{it} > A^r_{it}$ where the threshold parameters are linear functions of design aspects: $A^r_{it} = \boldsymbol{T}_{it}\boldsymbol{\alpha}^r$.

4.1 Gender and the income-wellbeing relation

A common finding in the literature on job satisfaction is that the pecuniary aspects of a job are less important to women than to men (see, for example, Booth and van Ours 2008). This was called into question by Conti and Pudney (2011), whose results suggested that responses from women interviewed F2F were subject to bias and that the gender difference largely disappeared when data from a more private PSC questionnaire were used instead. Table 5 explores this for satisfaction with income (waves 2-4) and job satisfaction (waves 1-4).

Table 5 Gender-income-design interactions in two satisfaction models

Coefficient	Satisfaction with income		Job satisfaction	
	GREO probit	FE regression	GREO probit	FE regression
Coefficients[1]				
Female	-0.283	-	0.418	-
	(0.250)	-	(0.372)	-
Income	0.629***	0.162	0.198*	0.070
	(0.062)	(0.115)	(0.119)	(0.185)
Female × Income	0.112	0.161	-0.117	0.004
	(0.083)	(0.157)	(0.157)	(0.232)
Female × F2F	0.859**	0.699	0.129	-0.135
	(0.380)	(0.461)	(0.417)	(0.519)
Income × F2F	0.079	0.030	-0.076	-0.241
	(0.096)	(0.115)	(0.127)	(0.154)
Female × Income× F2F	−0.259**	-0.181	-0.038	0.064
	(0.126)	(0.152)	(0.176)	(0.218)
Joint tests of design effects: P−values				
Additive design effects[2]	0.0000	0.0437	0.0000	0.000
F2F interactions	0.0687	0.1304	0.6545	0.2679

Standard errors in parentheses. Significance: * = 10%; ** = 5%; *** = 1%. [1] *Income* is log equivalised gross household income for the income satisfaction equation and log hourly earnings for job satisfaction. Other covariates included in the model are: age, age^2, single/widowed/divorced, no. children, non-white, wave dummies. [2] Design aspects in $\boldsymbol{T}_{it}$ are: CASI, CATI, Polar-labeled, 2-stage design, F2F.

In both the GREO and FE models, for both income and job satisfaction, the additive design variables are jointly significant at the 5% level. The FE regressions show no further design impacts and, indeed, no significant income effect or gender-income interaction at all. For the GREO models, there is some fairly weak evidence of a design interaction which could affect the empirical picture of gender differences in relation to money as a contributor to wellbeing, but only for the income satisfaction measure. In the GREO probit model for income satisfaction, the use of F2F rather than private interview modes seems to have two gender-specific effects: a large general increase in the levels of satisfaction reported by women; and a significant reduction in the female × income coefficient from 0.112 to -0.147. In other words, switching from private CASI to public F2F modes causes women significantly, on average, to downplay the importance of income in determining their income satisfaction. Both effects are individually significant at the 5% level, although jointly, the whole set of F2F-interactions are only significant at the 7% level. The result is consistent with Conti and

Pudney's (2011) findings for BHPS job satisfaction data, although the smaller sample sizes here reduce the statistical clarity somewhat.

4.2 Compensating variations for health conditions

Statistical models of wellbeing have often been used to estimate the income variation equivalent to events or resources like marriage, divorce, childbirth, unemployment and social capital (for example, Blanchflower and Oswald 2004, Di Tella and MacCulloch 2008 and Groot et al 2007). In health, the same approach has been used by Ferrer-i-Carbonell and van Praag (2002), Groot and Maassen van den Brink (2004, 2006), Mentzakis (2011), Zaidi and Burchardt (2005) and Morciano et al (2013) to estimate the personal costs of disease and disability. We have argued elsewhere (Hancock et al 2013) that this indirect method of contructing an estimate of the compensating variation (CV) as a by-product of a parametric model of wellbeing, is particularly sensitive to even minor misspecifications, often giving huge overestimates. Hancock et al (2013) argue for a more stable direct nonparametric approach, but indirect parametric estimation of the CV remains standard practice and so we examine the impact of survey design on it. We consider linear and quadratic models of overall life satisfaction, based on the latent regression (2), with the leading terms of the linear index specified as $\boldsymbol{x}_{it}\boldsymbol{\beta} = \beta_1 H_{1it} + \beta_2 H_{2it} + \phi(M_{it}) + ...$, where: H_{1it} is a binary indicator of the existence of a "long-standing health condition" that is not reported to cause any disability; H_{2it} indicates such a condition with associated disability; M_{it} is annual gross household income (in £'000) per equivalent adult; and $\phi(M_{it}) = \beta_3 M_{it}$ or $\beta_3 M_{it} + \beta_4 M_{it}^2$. In these two cases, the CV for health state $H_j (j = 1, 2)$ is $-\beta_j/\beta_3$ (linear model) or $-\left(B - \sqrt{B^2 - 4\beta_j \beta 4}\right)/2\beta_4$ (quadratic model), where $B = \beta_3 + 2\beta_4 M_{it}$.[3]

[3]Log income is often used in applied work, giving a CV of the form $M_{it} \exp\{-\beta_j/\beta_3\}$. This tends to produce even less robust CV estimates than the linear or quadratic income models and we do not report the results here.

Table 6 reports GREO probit estimates of the disability and income coefficients, and their interactions with the F2F interview mode. Again, additive design effects are highly significant, but here we are unable to detect any interaction between interview mode and health or income. Consistent with Hancock et al's (2013) findings, the implied CV estimates are extremely large, even for a non-disabling health condition: almost £28,000 for the linear model and around £19,000 at mean income for the better-fitting quadratic model. The F2F interaction raises these estimates still further, but the increase is not statistically significant.

Table 6 Gender-income-design interactions in two satisfaction models

	Linear in income	Quadratic in income
Coefficients[1]		
Income (£'000 p.a. per equivalent adult)	0.0078***	0.0182***
	(0.0016)	(0.0031)
Income^2		−0.0001***
		(0.00003)
Non-disabling health condition	−0.219***	−0.222***
	(0.068)	(0.068)
Disabling health condition	−0.461**	−0.453***
	(0.058)	(0.058)
Income × F2F	−0.0011	-0.0027
	(0.0027)	(0.0064)
Income^2 × F2F		0.00002
		(0.00008)
Non-disabling condition × F2F	−0.085	-0.085
	(0.117)	(0.117)
Disabling condition × F2F	−0.110	-0.119
	(0.092)	(0.092)
Joint tests of design effects: P−values		
Additive design effects[2]	0.0000	0.0000
F2F interactions	0.5785	0.7451
Estimated compensating variations ((£'000 p.a. per equivalent adult))[3]		
Non-disabling condition (not F2F)	27.95***	18.54**
	(10.58)	(7.65)
Non-disabling condition (F2F)	34.19**	20.92**
	(14.95)	(9.69)
$P-$ value for difference	0.718	0.706
Disabling condition (not F2F)	58.80***	53.21
	(14.96)	(34.0)
Disabling condition (F2F)	64.22***	57.61
	(21.73)	(56.4)
$P-$ value for difference	0.821	0.938

[1] Other covariates included in the model are: age, age^2, single/widowed/divorced, no. children, non-white, retired, wave dummies.
[2] Design aspects in $\boldsymbol{\xi}_{it}$ are: CASI, CATI, Polar-labeled, 2-stage design, F2F. [3] CV estimates at mean income for the quadratic model.

5 Survey design and health measurement

Health is a particularly important component of wellbeing and health measurement poses special difficulties. Our concern here is not for measurement of health itself, but whether the answers to important research questions about health may be affected by survey design.

We focus on two specific questions: (1) How large are gender differences in self-assessed health? (2) What is the magnitude of the income gradient in health? There are large applied literatures on both issues. Some authors have reported evidence of gender differences in health (Crimmins et al 2011) and there is a long-standing debate about the reality of these differences and their possible origin (Verbrugge 1980). The literature on the socioeconomic gradient of health is vast (World Health Organization 2008) and various measures of socioeconomic status have been used, notably education (Conti et al 2010) and income (Jones and Wildman 2008).

A large body of research relies primarily on self-reported subjective assessments of health or disability, which have been found to have significant predictive power for eventual mortality (Idler and Benyamini 1997, Doorslaer and Gerdtham 2003). Much of the empirical evidence used in economic evaluation of medical technologies and interventions rests on self-assessed health status through questionnaire instruments like EQ-5D (Herdman et al 2011) and the HUI (Horsman et al 2003), which are used to construct measures of Quality-Adjusted Life Years. These important questionnaire instruments are subject to much the same concerns about response behaviour as satisfaction assessments. The IP does not include EQ-5D or HUI, but there are several widely-used subjective health indicators. We focus on five, denoted $Y_1 ... Y_5$, which are all measured in waves 2-4 of the IP by the following questions:[4]

$\boldsymbol{Y}_1$: *Do you have any long-standing physical or mental impairment, illness or disability? By long-standing I 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.* At waves 2, 3 and 4, random subsets of respondents were assigned a shorter question wording: *Have you been, or are you likely to be, troubled for at least 12 months by any physical or mental impairment, illness or disability?*

[4]Where necessary, responses are recoded to be increasing with better health.

$\boldsymbol{Y}_2$: [For] *moderate activities, such as moving a table, pushing a vacuum cleaner, bowling or playing golf: does your health now limit you a lot, limit you a little, or not limit you at all?*

$\boldsymbol{Y}_3$: *Has a doctor or other health professional ever told you that you have any of the conditions listed on the card? [list of 17 disorders + "other"]* This is asked once only: at wave 2 or in a later wave on subsequent panel entry; responses are carried forward to later waves. We convert these responses to a single count of 0, 1, 2 or 3 or more diagnosed conditions from the list.

$\boldsymbol{Y}_4$: *In general, would you say your health is: 1 excellent; 2 very good; 3 good; 4 fair; 5 poor.*

$\boldsymbol{Y}_5$: *How dissatisfied or satisfied are you with* [...] *your health.*

The experimental treatments were as follows. At wave 2, a random subset of respondents were asked all questions by CATI. Of the others, $Y_1 ... Y_4$ came from F2F interview and the treatments administered for Y_5 were as set out in Table 1. At wave 3, all respondents were interviewed F2F for Y_1, Y_2 and Y_4; treatments for the satisfaction question Y_5 were as set out in Table 1. At wave 4, there was a random split between CASI and PSC, affecting Y_1, Y_2, Y_4 and Y_5. Indicator Y_3 was carried forward from responses at wave 2 (which had a CATI/F2F split), but later panel entrants all received the question F2F.

Analysis of health or disability can be carried out using a single health indicator or alternatively using a multiple indicator latent variable (MILV) approach. Both are common in the applied literature, and we consider the impact of survey design on both of them. Define $\boldsymbol{X}_i$ to be the vector of covariates believed to influence health; a particular element, say x_{i1}, is a covariate of particular interest; $\boldsymbol{T}_{ij}$ is a vector of dummy variables defining specific aspects of the experimental treatments relevant to indicator j. These design aspects are assumed to be additive. The single indicator models are again specified as the GREO probit form (2), with covariates $\boldsymbol{X}$ including a quadratic in age, gender, household size, number of children, marital status, educational qualifications, minority ethnic group, immigrant status, home ownership, log equivalised gross income,[5] the individual's share of household income, and dummies for managerial/ professional and unskilled status. The models are estimated separately for each wave 2..4. The stability of their design-specific thresholds is of no interest since they are typically treated as nuisance parameters in practice. Table 7 gives P-values for likelihood ratio tests of the hypothesis $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ for each indicator.

Table 7 Design impacts on 1-equation health models
(Likelihood ratio P-values in generalised ordered probit models)

	Wave 2	Wave 3	Wave 4
Y_1: long-standing impairment, illness or disability			
Additive design effect	0.000	0.000	0.000
Interaction: design × gender	0.512	0.550	0.433
Interaction: design × ln income	0.742	0.836	0.054
Y_2: health limits activities			
Additive design effect	0.021	-	0.000
Interaction: design × gender	0.039	-	0.077
Interaction: design × ln income	0.119	-	0.009
Y_3: diagnosed conditions			
Additive design effect	0.002	0.000	0.000
Interaction: design × gender	0.296	0.685	0.006
Interaction: design × ln income	0.167	0.231	0.000
Y_4: general health			
Additive design effect	0.624	-	0.000
Interaction: design × gender	0.335	-	0.352
Interaction: design × ln income	0.625	-	0.066
Y_5: satisfaction with health			
Additive design effect	0.000	0.000	0.000
Interaction: design × gender	0.067	0.420	0.244
Interaction: design × ln income	0.067	0.696	0.078

[1] see question wordings above; [2] baseline: F2F interview; [3] baseline: PSC; [4] baseline: open interview with late placement of satisfaction question (additive effects include also the effects of early placement, polar labeling, and 2-stage question structure). [5] Wald test P-value for interaction with CASI, CATI and wording treatments

The MILV model is attractive because the use of multiple indicators might be expected to give better robustness to measurement problems affectibg specific indicators. It consists of the threshold mechanism for each indicator, and a latent variable structure for latent health

[5]There is no household income variable available on a consistent basis for IP waves 1-4; our method of constructing income (including imputation of missing income components) is described in Holford and Pudney (2013).

η:

$$Y^*_{ij} = b_{ij}\eta_i + \upsilon_{ij} \tag{3}$$

$$\eta_i = \boldsymbol{X}_i\boldsymbol{\beta} + [x_{1i}\boldsymbol{\zeta}_i]\,\boldsymbol{\gamma} + u_i \tag{4}$$

$$b_{ij} = \boldsymbol{T}_{ij}\boldsymbol{\alpha}_j \tag{5}$$

where $\boldsymbol{\zeta}_i$ is again a set of variables representing specific design aspects, constructed from the collection $\{\boldsymbol{T}_{i1}...\boldsymbol{T}_{iJ}\}$. The random residuals $\varepsilon_{ij}, \upsilon_{ij}, u_i$ are iid $N(0,1)$ across individuals i and indicators j.[6] The two design variables in $\boldsymbol{A}_i$ are each constructed as the proportion of the five indicators $Y_{i1}...Y_{i5}$ for respondent i which are affected by specific design aspects: (i) telephone interviewing;[7] (ii) use of CASI for any question. The "deep" parameters in these models are the coefficient vectors $\boldsymbol{\beta}$ which (subject to scale normalisation) capture the influence of the covariates in $\boldsymbol{X}$ on latent health. Table 8 gives MILV estimates of the main effects for gender and log income, the corresponding interaction terms $\boldsymbol{\psi}$ and separate P-values for Wald tests of the elements of the hypothesis $\boldsymbol{\gamma} = \mathbf{0}$ corresponding to gender and log income.

Table 8 MILV model: Design impacts on the gender difference and income gradient in latent health.

	Wave 2		Wave 3		Wave 4	
	Gender[1]	Income[2]	Gender[1]	Income[2]	Gender[1]	Income[2]
Main effect (coefficient on female	-0.006	0.013	0.119	0.128*	-0.039	0.169***
gender or log income)	(0.121)	(0.082)	(0.097)	(0.065)	(0.086)	(0.057)
CATI	-0.010	0.283***	-0.449	0.011	-0.646	-0.011
	(0.138)	(0.085)	(0.515)	(0.134)	(0.427)	(0.113)
CASI	0.046	0.176	-0.091	0.105	0.433**	-0.043
	(0.982)	(0.256)	(0.619)	(0.158)	(0.173)	(0.049)
Wald test for coefficient shift[3] $\chi^2(2)$	0.014	11.097	0.799	0.469	9.832	0.760
(P-value)	0.993	0.004	0.671	0.791	0.007	0.684

[1] Coefficient of dummy variable for females. [2] Log gross annual income equivalised by modified OECD scale. [3] P-values for Wald tests of stability of the measurement threshold and factor loading parameters are all below 0.001.

[6] Evaluation of the likelihood was by 18-point Gauss-Hermite quadrature. We also attempted to fit a model with u distributed as a mixture of two normals (as used by Conti et al 2010); the ML estimator always gave a corner solution with a single error component.

[7] where used, CATI affects all five health indicators at wave 2 but, at waves 3 and 4, it can only affect indicator Y_3, if that derived from the wave 2 interview.

We confirm earlier evidence that measurement is strongly affected by survey design: with only one exception (the single-equation analysis of wave 2 general health indicator), parameter stability is strongly rejected for the ordered probit thresholds and MILV model factor loadings. Again, this suggests that comparison of average health states between groups interviewed in different ways is dangerous.

However, interaction tests show that the instability also affects deeper parameters. In the single-equation models (Table 7), there are two rejections at 5% for shifts in the estimated gender effect and income gradient (4 and 6 rejections respectively at the 10% level). In the MILV model (Table 8), we again reject stability. There are only two treatments that involve enough indicators and respondents to give good power for the tests: the CATI/F2F split at wave 2 and the CASI/PSC split at wave 4. At wave 2, estimates of both the gender difference and income gradient are small, positive and insignificant in the baseline F2F interview mode. The more private CASI and CATI modes both increase the income gradient, with the CATI effect highly significant. At wave 4, two relatively private modes of interview are compared: in the baseline PSC mode, there is a positive significant estimated income gradient and no significant gender difference. With CASI interviewing, the gender effect becomes significant and positive, which suggests that PSC questionnaires tend to produce more negative assessments, especially for women. This is consistent with Conti and Pudney's (2011) similar finding for PSC compared to F2F assessments of job satisfaction, but goes in the opposite direction to the more negative CASI assessments of life satisfaction that we found earlier. We lack a good theory of the way that interview mode interacts with income and gender as an influence on reporting behaviour for health, but there is clearly a need for such a theory.

6 Conclusions

There are three reasonably clear conclusions from our analysis of the wave 1-4 experiments in the UKHLS Innovation Panel, a couple of puzzling results, and some implications for the design of multi-wave experiments in large longitudinal surveys.

First, there is strong overall evidence that the choice of interview mode and question/response scale design has a detectable influence on the distribution of responses to questions on subjective health and wellbeing. This particularly true for computer-assissted self-interviewing (CASI) relative to other interview modes and there is some, weaker, evidence of an influence for the way the response scale is designed.

Second, the evidence for an influence of design features – especially interview mode – is stronger for female respondents than for males. This is consistent with evidence from other sources, and suggests a greater degree of sensitivity to the social context of the interview for women than men on average.

Our third conclusion is more important for the purposes of econometric analysis. We have taken three research questions as examples to assess the practical importance of these design effects: (*i*) Is there a gender difference in the impact of pecuniary factors on expressed wellbeing? (*ii*) What income variation is equivalent in wellbeing terms to a persistent health condition? (*iii*) How is the prevalence of ill-health related to demographic and socio-economic characteristics? We find that the answer to question (*i*) is influenced by the use of face-to-face (F2F) rather than more private modes of interview, with (after controlling for a wide range of other characteristics) women tending to give higher and less strongly income-related assessments of satisfaction with income, only when F2F interviewing is used. For research question (*ii*), we found no evidence for any effect of interview mode on the tradeoff between income and health, and therefore no impact on compensating income differentials. For research question (iii), we have detected a significant positive impact of the use of computer-assisted telephone interviewing (CATI) on the income gradient of latent health and that the gender difference in latent health appears to favour women in responses given via CASI but not via self-completion questionnaires. Despite the significant effects that we have found, on this evidence it seems fair to say that, with the possible exception of gender effects, the sort of conditional modelling used in economics seems more robust with respect to design differences than are simpler unconditional summary statistics.

But there are some puzzles accompanying these conclusions. At wave 3, which involved a more powerful comparison between fewer treatment groups, the evidence for design effects was actually weaker than at wave 2 – a finding which could possibly be explained in part by the 'contamination' of current responses by recalled past responses, as found by Pudney (2008, 2011). A second puzzle is that, at wave 4 where the comparison was between two relatively private interview modes (CASI and paper self-completion questionnaire), there was a large significant mean difference between responses, with CASI producing lower ratings of wellbeing. Given the similarity of the degree of privacy of those two modes, visual differences in response scale (e.g. vertical rather than horizontal presentation) may be involved in the impact that CASI appears to have.

Finally, resources like the UKHLS Innovation Panel are (arguably) a good way of ensuring that experiments are relevant to the reality of large-scale surveys but there is a risk that the resulting multiplicity of experiments within a moderately-sized sample may reduce power and complicate the interpretation of experimental effects, unless the complex of experiments can be designed in an integrated way. The problem of designing multiple experiments spanning multiple waves of a panel survey has not been studied systematically and it is not clear that the UKHLS Innovation Panel used in this paper has yet found a good way of managing the process of experimental design. Although randomised, the multi-treatment experiments considered here were confined to three or four waves and are arguably less effective in revealing framing and mode effects than the longer-term (and unplanned) BHPS experiment exploited by Conti and Pudney (2011), which involved sustained question repetition with different interview modes.

References

[1] Andrews F. M. and Withey, S. B. (1976). *Social Indicators of Wellbeing: Americans' Perceptions of Life Quality*. New York: Plenum Press.

[2] Aquilino, W. S. (1997). Privacy effects on self-reported drug use: interactions with survey mode and respondent characteristics. In Harrison L. and Hughes A. (eds.) *The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates*, 383-415, Rockville: National Institute on Drug Abuse, NIDA Research Monograph 167.

[3] Blanchflower, D. G. and Oswald, A. J. (2004). Well-being over time in Britain and the USA, *Journal of Public Economics* **88**, 1359-1386.

[4] Booth, A.L. and J.C. van Ours (2008). Job satisfaction and family happiness: the part-time work puzzle, *Economic Journal* **118**, F77-F99.

[5] Burton, J., Laurie, H. and Uhrig, S. C. N. (eds.) (2008). *Understanding Society* Some preliminary results from the wave 1 *Innovation Panel*. University of Essex: Understanding Society Working Paper no. 2008-03.

[6] Cameron, D. (2010). Speech on wellbeing, London 25 November 2010 http://www.number10.gov.uk/news/pm-speech-on-well-being/ (accessed 8 October 2013).

[7] Clark, A., Frijters, P. and Shields, M. (2008). Relative income, happiness and utility: an explanation for the Easterlin paradox and other puzzles, *Journal of Economic Literature* **46**, 95-144.

[8] Conti, G., Heckman, J. J. and Urzua, S. (2010). The Education-Health Gradient, *American Economic Review* **100**, 234-38.

[9] Conti, G. and Pudney, S. E. (2011). Survey design and the analysis of satisfaction, *Review of Economics and Statistics* **93**, 1087-1093.

[10] Crimmins, E. M., Ki, K. J. and Solé-Auró, A. (2011) Gender differences in health: results from SHARE, ELSA and HRS. *European Journal of Public Health* **21**, 81-91.

[11] Christian, L. M., Parsons, N. L. and Dillman, D. A. (2009). Designing scalar questions for web surveys, *Sociological Methods and Research* **37**, 393-425.

[12] Crossley, T. F. and Kennedy, S. (2002). The reliability of self-assessed health status, *Journal of Health Economics* **21**, 643-658.

[13] Deaton, A. S. (2010). Instruments, randomization and learning about development, *Journal of Economic Literature* **48**, 424-455.

[14] De Leeuw, E. (1992). Data quality in mail, telephone and face to face surveys

[15] Diener, E., Suh, E. M., Lucas, R. E. and Smith, H. L. (1999). Subjective well-being: three decades of progress, *Psychological Bulletin* **125**, 276-302.

[16] Ferrer-i-Carbonell, A. and van Praag, B. M. S. (2002). The subjective costs of health losses due to chronic diseases. An alternative model for monetary appraisal, *Health Economics* **11**, 709-722.

[17] Good, P. I. (2006). *Resampling Methods: A Practical Guide to Data Analysis* (3rd edition). Basel: Birkhäuser.

[18] Groot, W. and Maassen van den Brink., H. (2004). A direct method for estimating the compensating income variation for severe headache and migraine *Social Science and Medicine* **58**, 305-314.

[19] Groot, W. and Maassen van den Brink., H. (2006). The compensating income variation of cardiovascular disease, *Health Economics* **15**, 1143-1148.

[20] Groot, W. and Maassen van den Brink., H. (2007). Optimism, pessimism and the compensating income variation of cardiovascular disease: A two-tiered quality of life stochastic frontier model, *Social Science and Medicine* **65**, 1479-1489.

[21] Groot, W., Maassen van den Brink., H. and van Praag, B. M. S. (2007). The compensating income variation of social capital. University of Munich: CESIFO Working Paper no. 1889. *Health Economics* **11**, 709-722.

[22] Hancock, R. M., Morciano, M. and Pudney, S. E. (2013). Nonparametric estimation of a compensating variation: the cost of disability, University of Essex: ISER Working Paper 2013-26.

[23] Heckman, J. J., Stixrud, J. and Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics* **24**, 411482.

[24] Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. and Yavitz, A. (2010). Analyzing social experiments as implemented: a reexamination of the evidence from the HighScope Perry preschool program, *Quantitative Economics* **1**, 1-46.

[25] Helgeson, J.G. and Ursic, M.L. (1994). The role of affective and cognitive decision-making processes during questionnaire completion, *Public Opinion Quarterly*, 11(5), 493-510.

[26] Herdman M., Gudex C., Lloyd A., Janssen M., Kind P., Parkin D., Bonsel G. and Badia X. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L), *Quality of Life Research* **20**, 1727-1736.

[27] Herzog, A.R. and Bachman, J.G. (1981). Effects of questionnaire length on response quality, *Public Opinion Quarterly*, 45, 549-559.

[28] Holford, A. J. and Pudney, S. E. (2013). The *Understanding Society Innovation Panel*: Notes on the construction of a gross annual household income variable for waves 1-4. Mimeo, University of Essex.

[29] Horsman J, Furlong W, Feeny D, Torrance G. (2003). The Health Utilities Index (HUI): concepts, measurement properties and applications. *Health Quality Life Outcomes* **1**, 1-54.

[30] Hunter, J. (2005). Cognitive Test of the 2006 NRFU: Round 1. Washington DC: US Bureau of the Census, Study Series Report (Survey Methodology no.2005-07).

[31] Idler, E. L. and Benyamini, Y. (1997). Self-rated health and mortality: a review of twenty-seven community studies, *Journal of Health and Social Behavior* **38**, 21-37.

[32] Jenkins, C. R. and Dillman, D. A. (1997). Towards a theory of self-administered questionnaire design. In Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, K., Schwarz, N. and Trewin, D. (eds.) *Survey Measurement and Process Quality.* New York: Wiley.

[33] Jones, A. M. and Wildman, J. (2008). Health, income and relative deprivation: Evidence from the BHPS. *Journal of Health Economics* **27**, 308-324.

[34] Kahneman, D. and Krueger, A. B. (2006). Developments in the measurement of subjective wellbeing, *Journal of Economic Perspectives* **20**, 3-24.

[35] Kammann, N. R. and Flett, R. (1983). Affectometer 2: A scale to measure current level of general happiness. *Australian Journal of Psychology*, **35**, 259265.

[36] Kapteyn, A., Smith, J. P. and Van Soest, A. (2013). Are Americans Really Less Happy with Their Incomes? *Review of Income and Wealth*, **59**, 44-65.

[37] Kristensen, N. and N. Westergaard-Nielsen (2007). Reliability of job satisfaction measures, *Journal of Happiness Studies* **8**, 273-292.

[38] Krosnick, J. A., and Berent, Matthew K. (1993). Comparisons of party identification and policy preferences: the impact of survey question format, *American Journal of Political Science*, 37(3), 941964.

[39] Krueger, A.B. and D.A. Schkade (2008). The reliability of subjective well-being measures, *Journal of Public Economics*, 92(8-9), 1833-1845.

[40] McFall, S., Burton, J., Jäckle, A., Lynn, P. and Uhrig, C.S.N. (2013). Understanding Society: The UK Household Longitudinal Study Innovation Panel, Waves 1-5, User Manual. University of Essex: Institute for Social and Economic Research (https://www.understandingsociety.ac.uk/documentation/innovation-panel, accessed 30 Sep 2013).

[41] Morciano, M., Hancock, R. M. and Pudney, S. E. (2013). Disability costs and equivalence scales in the older population in Great Britain. *Review of Income and Wealth* forthcoming.

[42] Oswald, A. J. and Powdthavee, N. (2008). Does happiness adapt? A longitudinal study of disability with implications for economists and judges, *Journal of Public Economics* **92**, 1061-1077.

[43] Ralph, K., Palmer, K. and Olney, J. (2011). Subjective well-being: a qualitative investigation of subjective well-being questions. London: Office for National Statistics, research report.

[44] Schuman, H. and Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments in Question Form, Wording and Context*. New York: Academic Press.

[45] Schhwarz, N., Strack, F. and Mai, H. (1991). Assimilation and contrast effects in part-whole question sequences: a conversational logic analysis, *Public Opinion Quarterly*, 55, 3-23.

[46] Smith, T.W. (1979). Happiness: Time Trends, Seasonal Variations, Intersurvey Differences, and Other Mysteries, *Social Psychology Quarterly*, 42(1), 18-30.

[47] Stiglitz, J., Sen, A. and Fitoussi, J.-P. (2009). Report by the Commission on the Measurement of Economic Performance and Social Progress,

[48] Tourangeau, R., K.A. Rasinski and N. Bradburn (1991), "Measuring happiness in surveys: a test of the subtraction hypothesis", *Public Opinion Quarterly*, **55**, 255-266.

[49] Studer, R. (2012). Does it matter how happiness is measured? Evidence from a randomised controlled experiment, *Journal of Economic and Social Measurement* **37**, 317-336.

[50] Tourangeau, R. (1999). Context effects on answers to attitude questions, in Sirken, M. G., Herrmann, D. J., Schechter, S., Schwarz, N., Tanur, J. M. and Tourangeau, R. (eds.) *Cognition and Survey Research*. New York: Wiley.

[51] Van Doorslaer, E. and Gerdtham, U. G. (2003). Does inequality in self-assessed health predict inequality in survival by income? Evidence from Swedish data. *Social Science and Medicine* **57**, 16211629.

[52] Van Praag, B. and Ferrer-i-Carbonell, A. (2004). *Happiness Quantified. A Satisfaction Calculus Approach*. Oxford: Oxford University Press.

[53] Verbrugge, L. M. (1980) Sex differences in complaints and diagnoses, *Journal of Behavioural Medicine* **3**, 327-355.

[54] Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability, *Educational and Psychological Measurement*, 64, 956-972.

[55] World Health Organization (2008). *Closing the gap in a generation. Health equity through action on the social determinants of health*. Final report of the Commission on Social Determinants of Health. Geneva: WHO.

[56] Zaidi, A. and Burchardt, T. (2005). Comparing incomes when needs differ: equivalization for the extra costs of disability in the UK. *Review of Income and Wealth* **51**, 89-114.

Appendix: Additional tables

Table A1 Mean satisfaction scores and Kruskal-Wallis permutation tests

Treatment group	Place-ment	Women Health	Women Income	Women Leisure	Women Life	Women Job	Men Health	Men Income	Men Leisure	Men Life	Men Job
					Wave 2						
CASI full labels	Late	5.05	4.81	4.94	5.15***	5.11	5.53	4.58	4.56***	5.37*	4.97
CASI polar labels	Late	5.15	4.61	4.67**	5.20***	4.71**	5.05	5.07	5.05	5.40	5.48
Showcard full labels	Late	5.33	4.56	5.26	6.07	5.74	5.88	5.35	5.41	6.35***	5.35
F2F oral 2-stage	Late	6.00***	5.22	5.42	6.07*	5.46	5.13	4.56	5.48	5.48	5.35
Showcard polar labels	Late	4.52	5.30	5.04	5.43	5.38	5.32	4.64	4.14**	5.45	4.65**
Oral polar labels	Late	4.52*	4.77	4.92	5.26	4.90*	5.24	4.44	4.44	5.36	4.64*
CATI 2-stage	Late	5.28	4.92*	5.32**	5.75*	5.64***	5.19	4.54	5.13	5.54	5.45**
CATI polar labels	Late	5.27	4.73	4.89*	5.70	5.37	5.40	4.67	5.16	5.45	5.27
Showcard full labels	Early	5.52	5.33	5.41	6.07	.	5.14	4.87	5.30	5.57	.
F2F oral 2-stage	Early	4.96	5.08	5.20	5.72	.	5.76	5.14	5.19	6.24***	.
Showcard polar labels	Early	5.79**	4.25*	4.86	5.54	.	4.94	5.00	5.06	5.50	.
F2F oral polar labels	Early	4.91	4.27	4.41	5.18*	.	5.00	4.13	5.23	5.45	.
CATI: 2-stage	Early	5.10	4.66	5.19	5.82**	.	5.56	5.00	4.99	5.54	.
CATI: polar labels	Early	5.15	4.90	5.14	5.42	.	5.54	4.96	5.10	5.59	.
Overall mean		**5.21**	**4.78**	**5.06**	**5.51**	**5.40**	**5.35**	**4.74**	**5.05**	**5.53**	**5.26**
					Wave 3						
CASI full labels	Late	5.30	4.46**	4.97	5.50	5.06	4.92**	4.72	4.53**	5.40	4.81
CASI polar labels	Late	5.15*	4.96	4.91	5.49	5.16	5.38	4.94	4.97	5.41**	4.87
Showcard full labels	Late	5.29	4.88	5.03	5.75	5.51	5.59*	4.90	5.16	5.66	5.07
F2F oral 2-stage	Late	5.08	4.93	5.20	5.69	5.49	5.21	4.74	5.08	5.69	5.32*
Showcard polar labels	Late	5.35	4.88	4.97	5.61	5.24	5.36	4.93	5.14	5.67	5.27
F2F oral polar	Late	5.52**	4.96	5.21*	5.73	5.35	5.45	4.95	5.21	5.64	5.13
Overall mean		**5.29**	**4.88**	**5.05**	**5.63**	**5.30**	**5.36**	**4.88**	**5.07**	**5.60**	**5.07**
					Wave 4						
CASI full labels	Separate	4.67*	4.44*	4.67	5.16	.	4.70***	4.56	4.74	5.12	.
Paper self-completion	Separate	4.95*	4.64*	4.86	5.28	.	5.12***	4.71	4.85	5.24	.
Overall mean		**4.80**	**4.53**	**4.76**	**5.21**	**.**	**4.88**	**4.62**	**4.78**	**5.17**	**.**

Statistical significance in Kruskal-Wallis rank test (Monte Carlo permutation): *** = 1%; ** = 5%; * = 10%

Table A2: P-values for permutation tests on specific design aspects

Context	Wave	Women Health	Women Income	Women Leisure	Women Life	Women Job	Men Health	Men Income	Men Leisure	Men Life	Men Job
		Late *versus early questions*									
Showcard full labels	2	0.330	0.081	0.503	0.977	.	0.096	0.441	0.749	0.008	.
		0.744	*0.074*	*0.787*	*1.000*	.	*0.128*	*0.339*	*0.851*	*0.012*	.
Oral two-stage	2	0.002	0.710	0.489	0.194	.	0.266	0.177	0.699	0.116	.
		0.028	*0.772*	*0.620*	*0.216*	.	*0.245*	*0.285*	*0.611*	*0.079*	.
Showcard polar	2	0.026	0.865	0.813	0.757	.	0.555	0.575	0.149	0.990	.
		0.019	*0.940*	*0.757*	*0.809*	.	*0.465*	*0.570*	*0.149*	*1.000*	.
Oral polar	2	0.546	0.419	0.420	0.782	.	0.560	0.625	0.158	0.752	.
		0.416	*0.384*	*0.375*	*0.870*	.	*0.587*	*0.557*	*0.170*	*0.809*	.
CATI two-stage	2	0.581	0.157	0.718	0.408	.	0.085	0.071	1.000	1.000	
		0.410	*0.234*	*0.575*	*0.701*	.	*0.086*	*0.046*	*0.573*	*1.000*	.
CATI polar labels	2	0.663	0.531	0.131	0.128		0.491	0.154	0.863	0.270	
		0.492	*0.397*	*0.246*	*0.076*	.	*0.505*	*0.163*	*0.822*	*0.497*	.
		Full labels *versus polar labels*									
CASI	2	0.802	0.321	0.390	0.688	0.274	0.106	0.087	0.118	0.580	0.098
		0.732	*0.513*	*0.394*	*0.845*	*0.338*	*0.098*	*0.111*	*0.184*	*0.939*	*0.252*
CASI	3	0.335	0.035	0.957	0.541	.742	0.142	0.402	0.134	0.699	0.984
		0.532	*0.038*	*0.800*	*1.000*	*0.726*	*0.075*	*0.427*	*0.131*	*1.000*	*0.840*
Showcard	2	0.721	0.052	0.409	0.006	0.739	0.284	0.528	0.058	0.050	0.089
		0.550	*0.047*	*0.247*	*0.002*	*0.293*	*0.411*	*0.440*	*0.046*	*0.123*	*0.176*
Showcards	3	0.739	1.000	0.904	0.486	0.272	0.179	0.739	0.964	0.969	0.587
		0.727	*1.000*	*0.749*	*0.310*	*0.241*	*0.183*	*0.906*	*0.946*	*0.958*	*0.450*
		Two-stage *versus polar-labelled questions*									
Oral	2	0.008	0.174	0.161	0.032	0.093	0.089	0.123	0.158	0.029	0.115
		0.022	*0.065*	*0.079*	*0.007*	*0.226*	*0.370*	*0.134*	*0.171*	*0.132*	*0.162*
Oral	3	0.064	0.829	0.842	0.558	0.310	0.236	0.633	0.617	0.446	0.188
		0.026	*0.911*	*0.976*	*0.806*	*0.562*	*0.235*	*0.325*	*0.579*	*0.753*	*0.492*
CATI	2	0.363	0.447	0.039	0.027	0.011	0.532	1.000	0.964	0.504	0.129
		0.979	*0.741*	*0.032*	*0.121*	*0.001*	*0.340*	*0.611*	*0.730*	*0.714*	*0.296*

10,000 Monte Carlo replications. Roman type: H-statistic; Italic type: ANOVA

Table A2 continued: P-values for permutation tests on specific design aspects

		Women					Men				
Context	Wave	Health	Income	Leisure	Life	Job	Health	Income	Leisure	Life	Job
		CASI *versus PSC*									
Full labels	4	0.060	0.089	0.106	0.394	.	0.002	0.226	0.498	0.263	.
		0.011	*0.057*	*0.071*	*0.184*	.	*0.000*	*0.190*	*0.374*	*0.251*	.
		CASI *versus F2F*									
Full labels	3	0.917	0.063	0.564	0.098	0.203	0.009	0.417	0.021	0.354	0.351
		1.000	*0.083*	*0.816*	*0.146*	*0.072*	*0.006*	*0.490*	*0.027*	*0.234*	*0.344*
Polar labels	3	0.739	0.034	0.349	0.168	0.776	0.055	0.462	0.012	0.406	0.217
		0.554	*0.027*	*0.669*	*0.358*	*0.285*	*0.037*	*0.357*	*0.012*	*0.163*	*0.078*
		F2F (+ showcard) *versus CATI*									
Two-stage questions	2	0.180	0.306	0.900	0.798	0.750	0.478	0.650	0.420	0.076	0.747
		0.279	*0.214*	*0.895*	*0.477*	*0.594*	*0.661*	*0.613*	*0.400*	*0.192*	*0.762*
Polar labels	2	0.290	0.113	0.535	0.067	0.536	0.029	0.375	0.062	0.428	0.026
		0.125	*0.038*	*0.328*	*0.074*	*0.287*	*0.108*	*0.214*	*0.041*	*0.738*	*0.009*

10,000 Monte Carlo replications. Roman type: H-statistic; Italic type: ANOVA.

Table A3 Covariate sample means

Covariate	Mean	Covariate	Mean
Age	49.2	Log equivalised household income (£'000 p.a.)[a]	2.907
Single/widowed/divorced	0.189	Equivalised household income (£'000 p.a.)[a]	22.04
No. of dependent children	0.534	Weekly hours of work[b]	37.3
Non-white	0.086	Log Hourly wage (£)[b]	2.25
Retired	0.254	Hourly wage (£)[b]	11.07
		Non-disabling health condition	0.132
		Disabling health condition	0.216

[a] See Holford and Pudney (2013) for explanation of the method of constructing IP2 income variables; [b] Mean computed from positive sample values.

Values are pooled sample means for men and women and waves 1-4.