Understanding Society

Working Paper Series

No. 2015 – 05

September 2015

Survey response behaviour and the dynamics of self-reported health and disability: an experimental analysis

Annette Jäckle and Stephen Pudney

University of Essex

Understanding Society
THE UK HOUSEHOLD LONGITUDINAL STUDY

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Non-technical summary

A great deal of influential health research has been based on surveys asking respondents whether they are troubled by any long-standing illness, disability or infirmity. Such questions are also often used as a filter preceding a further question asking about difficulties with specific activities of daily life. Responses to this second question are widely used to construct measures of the extent or severity of disability, based on the number and types of difficulties a person reports.

Responses to the long-standing health question in panel surveys are very volatile: a large proportion of respondents reporting a problem in one interview no longer reported it in the following interview, and vice versa.

We report the results of an experiment carried in waves 6 and 7 of the *Understanding Society* Innovation Panel. The experiment was designed to identify reasons for the high rates of change in long-term illness/disability and to investigate whether using this question as a filter to determine who gets asked about difficulties with daily life impacts on estimates of the severity of disability.

Results suggest that the concept of a long-term health problem is not well defined. The question seems to measure better whether respondents have limitations caused by a health condition, than whether they have a condition at all: most respondents who stop reporting a long-term problem still have the condition, but it has improved, treatment or medication is more effective, or their activities have changed to make it less of a problem; among respondents who start reporting a problem, most claim that they had the condition at the previous interview. The results suggest at least 10 to 20% of reported health changes may contain errors. We examine the potential effects of such errors using both simulation methods and "correcting" errors by using the explanations given by respondents for apparent changes. Both methods show that the extent of errors observed can severely bias statistical analysis of the factors driving health changes.

The experiment also shows that if all respondents are asked about difficulties with daily life, more disability is reported than when the question is restricted to respondents who give an affirmative answer to a "filter" question asking about the existence of a long-term health condition. The latter is sometimes assumed to filter out respondents with only trivial disabilities, but we do not find this to be true. Instead, the filtered question design significantly reduces measures of onset and worsening of disability, and also leads to significant differences in estimates from multivariate models of disability.

Overall, our findings show that the design of survey questions can have a very important bearing on the evidence base that public policy on health and disability relies on. They also suggest that there may be a case for redesigning some of the questions used by many important surveys (including *Understanding Society*).

Survey response behaviour and the dynamics of self-reported health and disability: an experimental analysis

Annette Jäckle

Stephen Pudney

ISER, University of Essex

This version: September 21, 2015

Abstract: Disability research often uses survey questions asking whether respondents have a long-standing health problem, but longitudinal repetition of such questions produces implausibly high empirical transition rates into and out of ill-health. We exploit a repeated experiment in the *Understanding Society Innovation Panel* to identify reasons for these high transition rates, and to assess the common practice of using such questions to control who is asked further questions on difficulties with daily activities. Our results reveal ambiguity in the concept of a long-standing health problem and indicate significant biases in commonly-used measures and multivariate analyses of health dynamics and disability.

Keywords: Disability, Ill-health, Self-reported health, Measurement error

JEL codes: C2, C8, I10

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

This paper makes use of data from the Understanding Society Innovation Panel administered by ISER, University of Essex, and funded by the Economic and Social Research Council. (University of Essex. Institute for Social and Economic Research and TNS BMRB Waves 1-7, 2008-2014 [computer file]. 5th edition. Colchester, Essex: UK Data Archive [distributor], July 2015, SN: 6849, dx.doi.org/10.5255/UKDA-SN-6849-5). We are also grateful to the ESRC for financial support through the UK Longitudinal Studies Centre (award no. RES-586-47-0002). Participants in the 2015 Understanding Society conference made valuable comments.

1 Introduction

A great deal of influential health research has been based on surveys asking respondents whether they are troubled by any long-standing illness, disability or infirmity. Most major general population surveys carry such questions; the particular example we are concerned with here is carried in the UK *Understanding Society* panel survey and worded as follows: *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.* The resulting data are important as a focus of study in epidemiology (see, for example, Schroll et al. (1991); Idler and Benyamini (1997); van Doorslaer et al. (1997); Manor et al. (2001)). They are also widely used as explanatory variables in other fields, including labour economics (Disney et al. (2006), Garcia-Gomez et al. (2010)), social security and poverty analysis (Hancock and Pudney, 2013) and research on wellbeing (Dolan et al., 2008).

The detailed wording of self-report health questions varies from survey to survey – see Sturgis et al. (2001) for a review of UK data sources. There is UK evidence that cross-section analysis based on these questions are robust across surveys using different sample and question designs (Hancock et al., 2015). Self-reported health has also been shown to be predictive of future morbidity and mortality, even after conditioning on objective health measures (Idler and Benyamini (1997); Manor et al. (2001)). However, the picture is less reassuring if we look at repeated self-reported health measures longitudinally. As we show in section 2, there is a surprisingly high rate of transition in reported "long standing" health states, which suggests that there might be a degree of spurious churning in the data.

Questions about the existence of a substantial health condition are of interest in themselves, but they are also often used as a prefilter to a further question about difficulties with activities of daily life (ADLs).[1] Responses to the second stage ADL question are widely used to construct empirical measures of the extent or severity of disability, based on the number and types of difficulties that a person reports (see, for example, Zaidi and Burchardt (2005) and Morciano et al. (2014)). Influential policy-related examples include the Wanless (2006) Review and the Department of Health (2009) Green Paper on social care, both of

[1] UK examples include the *Understanding Society* panel and the Family Resources Survey (FRS).

which reached conclusions about the targeting of support for disabled people based on these measures.

There is reason to suspect that the use of filter questions may result in the total burden of disability being mis-measured. If the filter question is answered unreliably, measures of disability constructed from the reported difficulties with ADLs may be systematically biased. That bias could be positive if acquiescent behaviour causes respondents to favour the "yes" response to the filter question; or the bias may be negative, even with purely random response errors, because of the asymmetric question structure. A random false negative response bars entry to the ADL question and prevents recording any ADL difficulties, while a false positive does not necessarily lead to an offsetting over-estimate of ADL difficulties. Note that bias is not solely an issue of macro measurement: if response bias is related to the characteristics and circumstances of respondents, the inappropriate use of a prefilter to ADL questions could also generate incorrect distributional inferences: for example, on the targeting of public support for disabled people.

There is evidence in the generic survey methods literature that filtering has an impact on the eventual response, although most attention has been paid to the different setting of multiple filtered questions (Kreuter et al., 2011), where there is a possibility of respondents giving spurious "no" responses to avoid the burden of subsequent questions. The findings of Eckman et al. (2014) suggest that avoidance behaviour is more prevalent than acquiescence, indicating a tendency toward under-reporting. In the specific context of disability, Sweeney and Furphy (2008, pp. 228-229) also found some evidence from a Northern Ireland disability survey consistent with that view.

Any analysis of survey response behaviour is difficult in the absence of objective external validation data. While various biomarkers and formal diagnoses are available in some surveys, such measures have only an indirect connection with the main focus of interest, disability, and therefore cannot act as true validation data. An alternative approach to validation is to develop a better understanding of the way that respondents' behaviour may interact with survey design to generate inaccurate responses. In this study, we use randomised controlled experiments, implemented in waves 6 and 7 of the *Understanding Society* Innovation Panel (IP) to investigate the interaction of reporting behaviour and survey design, in relation to the reporting of change in health and disability.

We designed the experiment to do two things. First, reactive dependent interviewing (RDI) is used to give additional *ex post* information on the circumstances underlying any reported entry to or exit from a long-standing condition. Second, we use randomised differences in the administration of the filter question to identify the impact of filtering on measures of disability constructed from the ADL questions.

The paper is structured as follows. In section 2 we motivate the experiment by demonstrating the high empirical transition rates in health states in two major UK surveys, suggesting the strong possibility of measurement error in the form of spurious churning. We also use Monte Carlo simulation (in section 2.2) to demonstrate the magnitude of bias this could cause in estimation of typical statistical models. Section 3 describes our experimental design. Section 4 exploits randomised experimental treatments to provide empirical evidence on the nature and consequences of response error in a longitudinal context. We use RDI information about the circumstances relating to reported changes in health status to construct 'corrected' measures of alternative concepts of health and disability, demonstrating that summary measures of health dynamics and also the results from complex multivariate modelling can be sensitive to even low rates of reporting error. Section 5 presents experimental results on the use of a filter question controlling access to the ADL question from which our disability measures are derived, demonstrating the important impact of question design. Finally, section 6 suggests options for improving question design to produce more stable health measures.

2 Self-reported health and 'churning'

Our aim here is to document the surprisingly high rates of transition between health states which gave the original impetus for our experimental research. We look at the two longitudinal UK surveys that are most closely related to the Innovation Panel: the British Household Panel Survey (BHPS) and the main *Understanding Society* panel.

2.1 Health transitions in the BHPS and *Understanding Society*

The BHPS ran from 1991-2008 before being absorbed into *Understanding Society* from 2010 onwards. It used a number of health questions thet are rather different from those in *Under-*

standing Society; the closest parallel is the following question, carried in every wave except 1999 and 2004:

LIMITING: *Does your health in any way limit your daily activities compared to most people of your age?* [yes/no]

A further question, carried in all waves over 2002-18 uses a more subjective notion of disability:

DISABLED: *Can I check, do you consider yourself to be a disabled person?* [yes/no]

Another factual question was used from 1991 to 2000:

REGISTERED: *Can I check, are you registered as a disabled person, either with Social Services or with a green card?* [yes/no]

Let Y_t be the binary response to any of these question for respondents interviewed in year t. Table 1 summarises the prevalence, entry and exit rates for the three measures. Entry rates are estimates of $Pr(Y_{t+1} = 1|Y_t = 0)$, and exit rates are estimates of $Pr(Y_{t+1} = 0|Y_t = 1)$. We use survey weights to address nonresponse and other sources of bias and show separate estimates for working-age and older people respectively, where working age is defined as being over state pension age: 60 for women and 65 for men. Table 1 also gives the steady-state mean duration that would be implied by the empirical entry and exit rates in a simple Markov Chain equilibrium.[2]

For working-age people, the exit rate for the LIMITING indicator is slightly higher and the entry rate much lower than for older people, reflecting much greater prevalence in the older population but only moderately higher duration.

[2]If x is the exit rate, the equilibrium mean duration of completed poor health episodes is $(1 - x)/x$. This is intended only as a rough summary, representing the mean durations characterising an infinite-length realisation with constant transition rates and no absorbing state – assumptions which clearly do not apply to individual lives.

Table 1 Prevalence and transition rates: BHPS 1991-2008

	Working age			Above working age		
	LIMIT-ING	DIS-ABLED	REGIST-ERED	LIMIT-ING	DIS-ABLED	REGIST-ERED
Prevalence	0.14	0.07	0.04	0.33	0.25	0.13
Exit rate	0.28	0.29	0.18	0.22	0.20	0.22
Entry rate	0.05	0.02	0.01	0.12	0.08	0.04
Equilibrium mean duration (yrs)	2.6	2.5	4.4	3.5	3.9	3.6

More light can be shed on the dynamics of measurement by considering the overlap between entry and exit defined in terms of the LIMITING indicator and the alternative DISABLED and REGISTERED indicators. Table 2 looks at episodes of LIMITING entry and exit and shows the proportions of those cases which are also recorded as entry and exit by the DISABLED and REGISTERED criteria. The proportions are remarkably low – around one in ten concordance for the subjective DISABILITY indicator and less than one in twenty for disability registration.

Table 2 Correspondence between alternative transition indicators: BHPS 1991-2008

	Agreement with...	
	DISABLED	REGISTERED
Exit by LIMITING indicator	8.3%	2.2%
Number of cases	3,756	3,872
Entry by LIMITING indicator	11.4%	4.5%
Number of cases	4,300	4,862

The high-incidence, low-duration nature of the LIMITING indicator, and the minimal dynamic concordance between it and the DISABLED and REGISTERED indicators, gives some grounds for concern that it might be too volatile for the purposes of statistical modelling of health and disability, particularly in a dynamic context.

The main *Understanding Society* sample began in 2009 with roughly three times the sample size of the BHPS (approximately 30,000 households). Interviewing proceeds continuously through the year with each household interviewed on a 12-monthly cycle, but each wave takes two years to complete and thus overlaps with the preceding and succeding waves.

The survey has carried the limiting long-standing condition question from wave 2 (2010-11) onwards. Data up to wave 4 are available to us, so responses cover the calendar years 2010-2013. In addition to the main sample, *Understanding Society* has a separate panel of approximately 1,500 households, known as the Innovation Panel (IP), reserved for controlled experiments in survey design. The IP has annual interviews, conducted in the Spring of each year since 2008. Our experiment was implemented at waves 6 and 7 and is described in section 3, but the IP also carried the same question as mainstage *Understanding Society* for waves 2-5. Table 3 summarises the pooled-sample entry and exit rates for mainstage *Understanding Society* and the IP, by age group.

Table 3 Prevalence and transition rates in *Understanding Society* (main sample waves 2-4 and IP waves 2-5)

	Working age		Over working age	
	Main sample	IP	Main sample	IP
Prevalence	0.29	0.29	0.58	0.54
Exit rate	0.28	0.33	0.20	0.25
Entry rate	0.11	0.16	0.27	0.29
Equilibrium mean duration (yrs)	2.5	2.0	4.0	3.0

Table 3 reveals some differences with the BHPS LIMITING indicator. Entry rates are considerably higher for *Understanding Society*, while exit rates are broadly comparable. In particular, entry rates are high and considerably greater than exit rates for older respondents, while exit rates are larger than entry rates for working-age respondents. This corresponds better to the expectation that health problems tend to be of longer duration for older people. The most plausible explanation for the difference in entry rates between the BHPS and *Understanding Society* is that the BHPS question is age-referenced ("*compared to most people of your age*"), while the *Understanding Society* question is not.

2.2 The scope for bias: Monte Carlo simulations

The high empirical rates of exit from the ill-health state are hard to reconcile with the "long-standing" qualifier used in the question wording. It may be that the data are contaminated by spurious transitions which are the result of occasional error in the answers given by respondents, or the recording of those answers by interviewers. How important could such

errors be? We investigate this using Monte Carlo simulation of their impact on estimates of two representative statistical models, using three alternative assumptions about the measurement error process. Let Y_t be the 'true' health state and Y_t^* be the observed, possibly misreported/misrecorded state. Our three alternative assumptions are:

$$\text{Random misclassification:} \qquad Pr\left(Y_t^* \neq Y_t | Y_t\right) = \rho \tag{1}$$

$$\text{Reporting inertia:} \qquad Pr\left(Y_t^* = Y_{t-1} | Y_t \neq Y_{t-1}\right) = \rho \tag{2}$$

$$\text{Reporting fatigue:} \qquad Pr\left(Y_t^* = 0 | Y_{t-1} = Y_t = 1\right) = \rho \tag{3}$$

The first of these captures the idea of occasional, completely unsystematic, reporting errors. The second allows the possibility that there might be some delay in reporting a new condition, for example because it is not yet clear whether it can be expected to be long-standing. The third assumption captures the idea that, once a condition is under way (and would have been reported at onset), the respondent may not feel it necessary to report it again in the current year. In all of these cases, ρ represents a misreporting probability fixed at some *a priori* level as part of the simulation design.

We show the scope for measurement error by simulating standard estimators applied to two different econometric models.[3] The first is a simple regression model for life satisfaction (S_{it}), with the self-report of a limiting health condition (Y_{it}) used as a covariate together with log equivalised net income (M_{it}) and a vector of other covariates (X_{it}):

$$S_{it} = \beta Y_{it} + M_{it}\gamma + X_{it}\delta + u_i + \varepsilon_{it} \tag{4}$$

The second is a dynamic state-dependence (SD) model for Y_{it}:

$$Y_{it} = \mathbb{1}\Big(\beta Y_{it-1} + M_{it}\gamma + X_{it}\delta + u_i + \varepsilon_{it} > 0 \Big) \tag{5}$$

where ε_{it} is a random residual (normalised to have unit variance in (5)) and $\mathbb{1}(.)$ is the indicator function.

Estimation of model (4) is by fixed-effects regression, making no assumptions about the properties of the unobserved individual effect u_i, while the estimator for model (5) is random-effects probit, treating u_i as a Gaussian random variable independent of all covariates, and using the Wooldridge (2005) initial conditions approximation.

[3] See Hausman et al. (1998) for analogous simulation evidence on misclassification error bias in simple binary response models.

To make the simulation experiment as relevant as possible to real applications, we use actual BHPS data on the covariates X_{it}, which include six time-invariant variables representing ethnicity, gender and education, and ten time-varying covariates representing age, employment status, housing tenure and household size and structure. The X_{it} come from waves 15-18 of the BHPS, yielding 42,799 observations from 13,288 individuals.[4] The data on X_{it} are held fixed across the 500 Monte Carlo replications. The parameters of the experimental data generation process[5] are fixed at the values resulting from estimation of the model using actual BHPS data for Y_{it}. Full details of the simulation algorithm are given in the online appendix at http://iserwww.essex.ac.uk/home/spudney/?page_id=122.

Table 4 gives the mean and standard deviation across replications of the estimates of the health impact β in model (4) and the SD coefficient β and income coefficient γ in model (5).

[4] A different version of the health module was used at wave 14, so we use the last four waves only to avoid gaps or discontinuities in the series.

[5] The parameters are $\beta, \gamma, \delta, \sigma_\varepsilon^2$ and $u_i, i = 1...n$ for model (4), and $\beta, \gamma, \delta, \sigma_u^2$ for model (5).

Table 4 Monte Carlo simulation: impact of response error

Type of response error	Coefficient	Error rate ρ 5%	10%	20%
FE life satisfaction model: health coefficient $\beta = -0.222$				
Random misclassification	β	-0.131	-0.087	-0.047
		(0.016)	(0.014)	(0.011)
Reporting inertia	β	-0.228	-0.222	-0.208
		(0.020)	(0.018)	(0.019)
Reporting fatigue	β	-0.218	-0.215	-0.186
		(0.019)	(0.018)	(0.018)
Dynamic RE probit health model: SD coefficient $\beta = 0.351$*, income coefficient* $\gamma = -0.102$				
Random misclassification	β	0.189	0.100	0.031
		(0.040)	(0.033)	(0.025)
Reporting inertia	β	0.543	0.721	1.070
		(0.056)	(0.059)	(0.057)
Reporting fatigue	β	0.252	0.159	0.003
		(0.054)	(0.053)	(0.051)
Random misclassification	γ	-0.056	-0.036	-0.020
		(0.024)	(0.022)	(0.017)
Reporting inertia	γ	-0.098	-0.090	-0.080
		(0.031)	(0.032)	(0.028)
Reporting fatigue	γ	-0.097	-0.090	-0.086
		(0.031)	(0.032)	(0.028)

Sample size: $nT = 42,799$, $n = 13,288$; 500 Monte Carlo replications

The simulation results reveal considerable scope for measurement error bias, in some cases even with a modest rate of response error. For the FE regression analysis of life satisfaction, the bias is only large for the simple random misclassification case, with 40% attenuation of the health coefficient at a 0.05 error rate and almost 80% attenuation at a 0.2 error rate. The large biases are a consequence of the fact that the within-group transform used by FE regression reduces the signal-noise ratio in a measurement error context, thus magnifying classical measurement error bias. The two dynamic reporting processes of reporting delay or fatigue generate less bias, since they limit response error to the relatively small number of episodes involving some change in health status.

For the dynamic SD model, the largest bias is in the SD coefficient β, which is affected quite differently by the three alternative assumptions about the reporting error process. Under random misclassification, there is a large attenuation bias ranging from 46% at $\rho = 0.05$

to 91% at $\rho = 0.20$. In this case, there is also a similar degree of attenuation in the income coefficient γ. Together, these mean that the equilibrium income-health gradient would be seriously under-estimated. For the two dynamic error processes, bias is largely confined to the SD coefficient β, with the income coefficient γ only slightly attenuated. Reporting inertia and reporting fatigue have quite different implications for the estimated dynamic structure. Inertia generates a large positive bias in the SD coefficient, ranging from 55% to over 200% of the true coefficient at $\rho = 0.05$ and 0.20. In contrast, reporting fatigue attenuates β by 28% at $\rho = 0.05$ and almost 100% at $\rho = 0.20$. In both cases, the biases are large enough to give seriously misleading estimates of the equilibrium income-health gradient.

The very different character of measurement error biases generated by different (plausible) patterns of reporting error show how important it is to understand reporting behaviour in relation to health and disability. The experiments set out and analysed in sections 3-5 are intended to contribute to that improved understanding.

3 The Innovation Panel experiments

Our experiments were implemented in waves 6 and 7 of the IP and had two separate strands: (*i*) reactive dependent interviewing (RDI) was used to investigate the factors underlying change in self-reported measures of long-standing illness or disability, using responses collected at wave 5 as the initial reference point; (*ii*) randomised controlled trials of three variants of the health-disability instrument were used to show how question design interacts with response behaviour to generate empirical measures of disability. See Jäckle et al. (2014) for further details of the experiment and others carried in waves 1-7.

The experiment is based on the following four questions.

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]

RDI: *Just to check, our records show that last time when we interviewed you on [date], [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]?* [1 There

is an error in the records; 2 I [still have/had] the same health condition but it is [not as bad/worse] now; 3 I [still have/had] the same health condition but treatment or medication is [effective/less effective] now; 4 The condition is much the same as last year, but my activities have changed, so it is [less/more] of a problem now; 5 [I no longer have this/This is a new] health condition; 6 Other reason]

FREETEXT (if RDI = 1 or 6): *Please explain the [error / other reason for the difference]*

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]

Sample members were randomly (by household) allocated to one of three experimental groups; group A covered approximately half the sample and groups B and C a quarter each. The details of the treatment received by each group is set out in Table 5. Members of treatment group A all received the HEALTH question and (about ten minutes later) the unfiltered ADL question. Group B were not asked the HEALTH question and, instead, received the ADL question on its own with no filter. Respondents in group C received the standard *Understanding Society* instrument: they were asked the HEALTH question and then received ADL only if a health condition was declared. Thus there are three sources of difference here: whether or not HEALTH was encountered before ADL (group B *vs.* groups A and C); the time gap between the HEALTH and ADL questions (group A *vs.* group C); and the use of a filter to control access to the ADL question (group C *vs.* groups A and B). Every individual received exactly the same treatment at wave 7 as (s)he did at wave 6. The dataset we work with includes all individuals interviewed at both wave 5 and 6 and all who were interviwed at waves 6 and 7. Anyone providing an interview at waves 5 and 7 but not wave 6 is excluded. The sample numbers given in Table 5 may vary slightly from the numbers involved in particular comparisons because of a very small amount of item non-response.

Table 5 Experimental design

Group	Sample numbers at wave...[1] 5	6	7	Approx. 5 mins. into interview RDI + HEALTH	FREETEXT	ADL[2]	Mid-interview ADL[2]
A	-	865	766	✓	✓ (if reported change)	-	✓
B	-	429	378	-	-	✓	-
C	1,991	409	359	✓	-	✓ (if HEALTH=Yes)	-

[1] Counts exclude those who did not respond to the HEALTH question (or ADL, for group B). [2] The ADL question was not asked in proxy or youth interviews.

4 Dynamic response error

1,258 respondents from groups A and C answered the HEALTH question at both wave 5 and wave 6, a third (418) reporting an initial health condition or disability at wave 5, of whom 78 reported no condition at wave 6: an exit rate of 19%. Among the 840 respondents at wave 5 who reported no long-standing condition, 120 reported such a condition at wave 6: an entry rate of 14%. Between waves 6 and 7, the exit and entry rates were 25% and 7% respectively. These transition rates are slightly lower than the corresponding rates in the main *Understanding Society* sample and earlier IP waves, but they remain surprisingly high in view of the "long-standing" qualifier used in the question wording.

4.1 Alternative health concepts and 'corrected' measures

It is not possible to discuss measurement error without a clear definition of the concept that is to be measured, and there is no guarantee that the analyst's concept of a long-standing health condition coincides with the concept used by respondents when forming their answers. There are two classes of response problem: *pure measurement error*, where the analyst and respondent share a common understanding of the relevant concept, but the response is given or recorded incorrectly; or *conceptual mismatch*, where the response is correct in its own terms but is based on a conceptual interpretation that differs from the analyst's.

The RDI responses summarised in Table 6 suggest that both pure measurement error

and conceptual mismatch are present.[6] There is a non-negligible number of cases where respondents give a definite indication of an error in the records: 6% and 8% for reported exits and 21% and 10% for reported entries, for 2006 and 2007 respectively. If applicable to general population samples, the Monte Carlo simulation results of section 2.2 suggest that these error rates could be sufficient to generate large biases in the coefficients of conventional econometric models involving self-reported health variables.

Table 6 Reasons for changes in long-term health status

		No. exits		No. entries	
R_{it}	Reported reasons for change	wave 6	wave 7	wave 6	wave 7
1	Error in the records	4	7	20	3
2	Condition improved/worsened	9	16	9	6
3	Treatment effectiveness changed	15	17	3	1
4	Activities changed	6	5	3	1
5	Condition started/ceased	11	5	29	12
6	Other reason	1	2	5	3
0	No explanation given	1	0	1	1
Number of group A exits/entries		46	49	69	27

Table 6 also reveals some ambiguity about the concept of health involved and potential for conceptual mismatch. It is possible to discern five separate issues arising in the standard list and free-text explanations: (*i*) the existence of a definite condition; (*ii*) the severity of that condition; (*iii*) the effect of medical treatment in alleviating its effects; (*iv*) the pattern of normal activity that the person chooses; (*v*) the resulting limitations on that pattern of activity.

The size of our experimental samples makes it infeasible to pursue all of these elements separately, but the wording of question HEALTH suggests two broad concepts which we attempt to distinguish. One is *existence of a condition* (C_{it}): "Do you have any long-standing physical or mental impairment, illness or disability?" The other is the *limitation of a condition* (L_{it}): "anything that has troubled you". The latter is more complex than the former, since "trouble" will be affected by severity of the condition, medical treatment and desired pattern of activity. For experimental group A, when a respondent reports a

[6]Note that a few respondents gave two reasons for the reported change in health state and some gave no explanation, so the reasons cited do not sum to the total number of reported transitions. There are no instances of RDI at wave 7 suggesting an error at wave 6 when RDI had also been used at wave 6, so it is possible for us to use the sequence of RDI outcomes without any need to resolve conflicts between them.

changed health state, the RDI question tells us more about the nature of the change. We use this information (and, where relevant, the free-text explanation) to construct alternative 'corrected' indicators of the health state corresponding to these two distinct health/disability concepts. It should be borne in mind that only partial data adjustment is possible because we do not have RDI follow-up for people who did not report any change, so a reporting error that persists for three periods would remain uncorrected.

We construct the indicator C_{it} for the weaker concept of a health condition primarily by reversing apparent exits or entries (setting $C_{it} = C_{it-1} = 1$) whenever RDI suggests that the same condition existed in both periods ($2 \leq R_{it} \leq 4$). The narrower limitations indicator L_{it} makes the same substitution only when RDI suggest some conflict with normal activity ($R_{it} = 4$). The editing process is in fact rather more complex than this, since it also takes account of the free-text explanations available in some cases. The number of corrected data points is small in total – around 4% of the group A observations in waves 6 and 7 for C_{it} and 1% for L_{it}, but large relative to the number of reported exit transitions – 67% of exits for C_{it} and 12% for L_{it}. The appendix gives a detailed specification of the two data editing algorithms used to construct C_{it} and L_{it}, and the code which implements it is included in the online appendix at `http://iserwww.essex.ac.uk/home/spudney/?page_id=122`.

The effect of data adjustment on transition rates varies between the two health concepts. The entry and exit rates for initially reported ill-health are 8.9% and 16.7% respectively for teatment group A. For the C_{it} indicator, these are reduced to 8.3% and 7.0% and, for the L_{it} indicator, they are changed less, at 8.6% and 13.9%. Thus the RDI evidence suggests that the main reason for the high exit rate is not full recovery from a health condition, but partial improvement over time (often as a result of treatment) or the intermittent nature of conditions like asthma and migraine. A lot depends, therefore, on whether one is interested in the prevalence of disease or the impact that disease has on everyday life. Measurement problems appear to be a bigger obstacle to analysis of the former than of the latter.

The Monte Carlo experiment in section 2.2 revealed quite different consequences of two hypothetical dynamic misreporting processes: inertia and fatigue. The former may occur when a transition takes place, the latter when an adverse health state persists across two periods. Using the C_{it} indicator, simple inertia and fatigue error rates can be constructed as $Pr(C_{it} \neq Y^*_{it} | C_{it} \neq C_{it-1})$ and $Pr(C_{it} \neq Y^*_{it} | C_{it-1} = C_{it} = 1)$. Their empirical counterparts are

1/214 = 0.5% and 59/492 = 12.0% respectively; the analogous rates for the L_{it} indicator are 0.4% and 2.5%. These rates are only indicative: they are based on small sample numbers and rely on a partial and somewhat arbitrary error correction process, but they suggest that the main area for concern is the possibility of response fatigue rather than inertia, and primarily in relation to the existence, rather than limiting nature, of a health condition.

This is confirmed by logit estimates summarised in Table 7. We construct a binary indicator of response error as $E_{it}^C = \mathbb{1}(C_{it} \neq Y_{it}^*)$ and binary covariates identifying the situations in which fatigue and inertia can occur: $F_{it}^C = \mathbb{1}(C_{it-1} = C_{it} = 1)$ and $I_{it}^C = \mathbb{1}(C_{it-1} \neq C_{it})$, with variables $E_{it}^L, F_{it}^L, I_{it}^L$ constructed analogously for the L_{it} health concept. We then estimate logit regressions for the probability of error, $Pr(E_{it} = 1|F_{it}, I_{it})$. For both the C_{it} and L_{it} health concepts, the influence of fatigue is larger than the insignificant estimated effect of inertia, particularly so for the C_{it} definition of health. This points to a specific pattern of response behaviour which sometimes fails to distinguish existence from onset and may follow a line of reasoning by respondents: "I've told you about this illness before, so I don't need to tell you again".

Table 7 Logit models of the dynamic pattern of response error

	Health concept	
Logit coefficient	existence of condition: C_{it}	limitation from a condition: L_{it}
Inertia: I_{it}	0.367	0.218
	(1.158)	(1.108)
Fatigue: F_{it}	3.735***	2.079***
	(0.594)	(0.653)
Pseudo-R^2	0.224	0.083)

Standard errors in parentheses. Significance: * = 10%; ** = 5%; *** = 1%. $N = 1,631$.

4.2 Bias in multivariate models of health and life satisfaction

We examine the impact of our data 'correction' procedure on the estimation results for typical multivariate models, focusing on specific models of health dynamics and of the influence of health on life satisfaction. Both are estimated using data from the three waves 5-7 on individuals in treatment group A, for our three alternative measures of health: Y_{it}^*, C_{it} and I_{it}. Estimates of selected coefficients from these estimated models are presented in Table 8.

The model of health dynamics is the standard autoregressive panel data probit model:

$$Y_{it} = \mathbb{1}\left(\alpha Y_{it-1} + X_{it}\beta + u_i + \varepsilon_{it} > 0\right) \tag{6}$$

where the coefficient of interest α measures the persistence of ill-health, X_{it} is a vector of covariates, u_i is an unobserved random effect and ε_{it} is a white noise residual. The model is estimated using the Wooldridge (2005) method for dealing with initial conditions.

We find a large increase in the estimated degree of persistence (α) in the dynamic model (6), when we substitute the corrected indicator for existence of a health condition (C_{it}) for the initial self-report ($Y^{\star}_{it}$): existence of a condition in the previous wave is estimated to increase prevalence by 25 rather than 7 percentage points for an average individual. For the limiting health concept (L_{it}), error correction has a much smaller effect, raising prevalence by 9 percentage points rather than 7.

For life satisfaction S_{it}, measured on a 1-7 scale, FE regression has become the standard estimation approach in the happiness literature, but we were unable to detect any significant health effects using FE regression. This is a consequence of poor estimation precision, since the within-group transformation used to eliminate the u_i greatly depletes sample variation in this 3-wave panel. Instead, we use the following random effects model where the individual effect u_i is assumed independent of the covariates X_{it}:

$$S_{it} = \alpha Y_{it} + X_{it}\beta + u_i + \varepsilon_{it} \tag{7}$$

In this static context, adjustment for reporting errors in health transitions has very little effect on the estimates: substituting C_{it} or L_{it} for $Y^{\star}_{it}$ changes the health coefficient by a negligible amount. This is in line with the Monte Carlo simulation results in Table 4 and is a consequence of the small proportion of cases involving reported health transitions which are at risk of the error our experiment is designed to detect.

Table 8 Estimated models of health dynamics and life satisfaction: health coefficients

Coefficient	Health measure		
	Y^*	C	L
Dynamic health model			
$\hat{\alpha}$: lagged health	0.914***	2.807***	1.232***
	(0.327)	(0.394)	(0.346)
(Marginal effect at $u = 0$)	0.069	0.250	0.091
	(0.046)	(0.132)	(0.059)
Life satisfaction model			
$\hat{\alpha}$: health	−0.264***	−0.276***	−0.263***
	(0.091)	(0.090)	(0.091)

[1] Significance of estimated parameters: * = 10%; ** = 5%; *** = 1%

5 The effect of a filter question on disability analysis

We now examine the consequences of using the HEALTH question as a filter preceding the ADL question. We compare the responses from group B (who were asked the ADL question with no filter) with those from group C (who received the filtered version which is currently used in the main *Understanding Society* sample). Two indicators of disability are used to summarise the results: D_1 is a binary indicator taking the value 1 if one or more of the twelve ADL difficulties is declared by the respondent,[7] and D_2 is a count index of the number of ADL difficulties declared, taking the value 0 for the majority of the sample who report no ADL difficulty.

Table 9 gives empirical estimates of three summary measures: $E(D_1)$ is a measure of disability prevalence; $E(D_2|D_2 > 0)$ measures the average severity of disability among the disabled; and $E(D_2)$ measures the aggregate per capita burden of disability. Significance tests for between-group equality of means are based on the bootstrap clustered at the household level.

[7]Note that this is not necessarily identical to Y: some respondents indicate the existence of a health condition without indicating any ADL difficulty.

Table 9 The effect of filtering on disability responses

Disability measure	Filter (group C)	No filter[1] (group B)
Wave 6		
Prevalence: $E(D_1)$	0.244	0.310**
Unconditional mean count: $E(D_2)$	0.548	0.716*
Conditional mean count: $E(D_2\|D_2>0)$	2.240	2.308
Change in prevalence: $E(\Delta D_1)$	0.032	0.091**
Change in mean count: $E(\Delta D_2)$	0.005	0.140**
Wave 7		
Prevalence: $E(D_1)$	0.203	0.304***
Unconditional mean count: $E(D_2)$	0.510	0.722*
Conditional mean count: $E(D_2\|D_2>0)$	2.507	2.374
Change in prevalence: $E(\Delta D_1)$	-0.041	0.012
Change in mean count: $E(\Delta D_2)$	-0.031	0.088*

[1] Significance of mean difference relative to group C: * = 10%; ** = 5%; *** = 1% (household-clustered bootstrap test, 1000 replications)

Three conclusions emerge. First, the use of a filter question reduces the measured prevalence of disability and (at the 10% significance level) also the per capita aggregate burden. The difference in mean prevalence between groups B and C is 7 and 10 percentage points at wave 6 and 7 respectively, equivalent to 21% and 33% in terms of the reduction in the aggregate number of people reporting any ADL difficulty. The reduction in the measured aggregate burden caused by filtering is of the same magnitude (23% and 29% respectively).

It is sometimes argued that the filter question is necessary to screen out trivial difficulties that do not warrant description as disability. If that were so, we would expect the mean disability count in the subset of people reporting disability ($E(D_2|D_2 > 0)$) to be lower in group B (no filter) than in group C (filter).[8] Instead, the estimated differences are modest in size (-3% and +5% in waves 6 and 7) and statistically insignificant ($P = 0.777$ and $P = 0.620$). We therefore find no evidence that the filter question is necessary to avoid spurious inclusion of trivial conditions in measured disability, so there is a concern that, unlike ELSA, surveys like *Understanding Society* and FRS, which use a filtered question structure, might miss some substantial disabilities.[9]

[8] The effect on the measured aggregate extent of disability ($E(D_2)$) then depends on the balance of these two, but we would expect it to be greater in group B, assuming that removal of the filter has no effect on the ADL responses of people who would satisfy the filter.

[9] Note that Hancock et al. (2015) find that ELSA generates higher empirical disability rates than FRS in comparable areas (*e.g.* incontinence).

A third conclusion from Table 9 concerns the pattern of within-individual change in disability over time. The use of a filter question significantly reduces empirical measures of the mean wave-to-wave change in D_1 and D_2. The tendency of filtering to eliminate evidence of the increase in incidence and severity of disability over time in a panel of individuals gives grounds for concern in relation to dynamic analysis of disability.

We also investigate the effect of treatment group on results from a multivariate statistical model of disability incidence in which the distribution of the count measure of disability, D_{2it}, is specified as zero-inflated negative binomial (ZINB) conditional on explanatory covariates X_{it} (Mullahy, 1986). The ZINB is a mixture model with two parts: a logit structure, $P_0(X_{it}) = (1 + e^{X_{it}\alpha})^{-1}$, gives the probability that disability is necessarily zero. With probability $1 - P_0(X_{it})$, the ADL count D_{2it} has a negative binomial distribution with location parameter specified as $\mu(X_{it}) = e^{X_{it}\beta}$. An appealing way to interpret this model is that part 1 of the mixture describes the population of people who are definitely non-disabled and would not consider reporting an ADL difficulty, while part 2 of the mixture uses the negative binomial distribution to describe the population of people with some degree of impairment who may or may not see it as sufficiently serious to be worth reporting at interview. Thus a report of $D_1 = D_2 = 0$ can occur in one of two ways - as straightforward report by a non-disabled person, or as a response by a person with some disability who feels at the time of interview that the ADL difficulties do not warrant reporting. $P_0(X_{it})$ is the predicted probability of the first type of zero and $1 - P_0(X_{it})\mu(X_{it}$ is the predicted mean disability burden $E(D_2)$.

We use a simplified set of covariates comprising age and three binary indicators of socioeconomic status: home-ownership, post-compulsory educational qualifications and managerial/professional/skilled non-manual occupation in current or most recent job. The model is estimated separately for treatment groups B and C. For each group the same two covariates, age and homeownership, are strongly significant, with people who are older or non-homeowners having a higher probability of being disabled and a higher expected ADL difficulty count if they are disabled. All other covariates are jointly insignificant.[10] Table 10 presents mean predictions and marginal effects computed from the estimated models.

[10] $\chi^2(6)$ Wald tests, $P = .68$ and $.56$ for groups B and C respectively.

Table 10 The effect of filtering on estimated ZINB models for treatment groups B and C

Disability measure	Filter (group C)	No filter (group B)
Mean predictions		
Mean $P_0(X_{it})$	0.664	0.472
Mean of $E[D_{2it}\|X_{it}]$	0.561	0.723
Wald test for parameter equality $\chi^2(19)$ =		32.9**
Marginal effects at covariate means		
Age	0.022	0.034*
	(0.002)	(0.02)
Homeowner	–0.570	–1.279*
	(0.098)	(0.147)

Household-clustered standard errors in parentheses; statistical significance of difference with group C: * = 10%; ** = 5%; *** = 1%

The estimated ZINB model shows significant structural instability across treatment groups, with a considerably higher average predicted disability burden and lower predicted prevalence of non-disability in the unfiltered responses from group B. The main source of the difference is in the coefficients of age and homeownership, with both having a stronger estimated association with disability when no filter is used. Thus the demographic and socioeconomic gradients in disability are attenuated by the use of a filtered question structure.

These results are potentially important for policy purposes. If no filter is used (group B), there is a higher projected volume of disability, and disability appears more strongly concentrated among people who are relatively old and in rented accommodation. If used for policy projections, a model based on an unfiltered survey question would therefore project higher social care needs and a higher proportion of disability among people without the means to meet the costs of their disability. This is a good illustration of the importance of apparently arcane technical design issues for practical policy analysis.

6 Conclusions

In this study, we have used randomized experiments involving reactive dependent interviewing (RDI) and comparison of alternative questionnaire designs to investigate the problems of survey response error in self-reported long-term limiting illness and associated disability

indicators. These are widely used measures available in many important survey datasets. The possibility of bias in the data and in research results based on them is an important practical issue.

We have reached conclusions in five areas. First – and possibly most important – concerns the ambiguity of the concept of long-term limiting illness (LLI). Our analysis suggests that self-reported LLI is a conceptually unreliable indicator of the existence of a long-term health condition. It is much more robust as an indicator of health-induced limitations on everyday activity but is prone to a surprising degree of 'churning' when observed longitudinally, arising from the inherently variable character of many illnesses. Respondents showed some confusion about what was required in cases where the impairment is variable, particularly for conditions like asthma, migraine and arthritis which may be intermittent and variable in intensity. The concept of a long-term health condition is ambiguous in these cases and short-term random factors (for example, the recency of an asthma attack) may generate reported health transitions even if the daily probability of an asthma attack has remained constant.

In addition to the conceptual ambiguity facing respondents and researchers, evidence from RDI suggests a significant rate of misreporting by respondents, or misrecording by interviewers, of change. Although the number of experimental cases is not large enough to permit estimation of a full model of reporting error, the dominant form of error appears to be response 'fatigue' leading to spurious reports of exit from LLI. Once the existence of a condition has been reported, there is a significant chance of the report not being repeated at the following wave, even if the health problem remains in place.

A third finding is that even modest rates of dynamic misreporting can have serious implications for econometric analysis. The main problem is in dynamic multivariate modelling of LLI, where Monte Carlo simulations and RDI results from our experiment both show large biases in estimated coefficients of dynamic probit models involving state dependence. The effect of measurement error is to bias downwards the state dependence effect and attenuate the estimated socioeconomic gradient in health. On the other hand, we find no evidence for serious bias in models where the LLI indicator is used as an explanatory variable in a panel data regression model of (for example) life satisfaction.

Our fourth conclusion relates to the construction and use of disability measures. Survey measures of disability are usually constructed from responses to questions about difficulties

with a standard list of specific activities of daily living (ADL). The ADL question is typically asked only of respondents who have reported the existence of a limiting health condition, so that the LLI question acts as a filter controlling access to the ADL question used to measure disability. An alternative unfiltered design simply asks the ADL question of all respondents, without precondition. We have found evidence that the prevalence, aggregate burden and change over time in disability are much larger when measured by the unfiltered questionnaire design rather than the standard filtered design. It is sometimes argued that the use of a filter qestion is a way of screening out trivial mild impairments but we have found no evidence to support that view – conditional on the reporting of any disability, there is no significant difference between the mean severity of disability reported by the groups randomly allocated the filtered and unfiltered designs. These differences also have an impact on statistical models of the demographic and socioeconomic gradients of disability, where we find significant coefficient differences in an illustrative count data model. The use of the filtered question design was found to reduce substantially and significantly the estimated magnitude of these gradients.

Finally, we can make two recommendations for improvement in survey design. The use of RDI follow-up to questions about long-term limiting illness can shed valuable light on the meaning of reported change and gives researchers a way of assessing the robustness of their findings. Although RDI increases survey costs by complicating computerised interview scripts and lengthening the interview, the cost increases are not large and, in our view, RDI follow-up should be considered as a standard feature for longitudinal LLI survey instrumentss. A bigger challenge for future work is the need to present respondents with a question design that gives a clearer conceptual basis for LLI, accommodating situations where there is a long-standing health condition which imposes serious but intermittent limitations on normal activities. In our view, the development of such a question design should be a priority for survey designers.

References

Department of Health (2009). *Shaping the Future of Care Together*. Department of Health, London.

Disney, R., Emmerson, C., and Wakefield, M. (2006). Ill health and retirement in Britain: A panel data-based analysis. *Journal of Health Economics*, 25:621–649.

Dolan, P., Peasgood, T., and White, M. (2008). Do we really know what makes us happy? a review of the economic literature on the factors associated with subjective well-being. *Journal of Economic Psychology*, 29:94–122.

Eckman, S., Kreuter, F., Kirchner, A., Jäckle, A., Tourangeau, R., and Presser, S. (2014). Assessing the mechanisms of misreporting to filter questions in surveys. *Public Opinion Quarterly*, 78(3):721–733.

Garcia-Gomez, P., Jones, A. M., and Rice, N. (2010). Health effects on labour market exits and entries. *Labour Economics*, 17:62–76.

Hancock, R. M., Morciano, M., Pudney, S. E., and Zantomio, F. (2015). Do household surveys give a coherent view of disability benefit targeting? a multisurvey latent variable analysis for the older population in Great Britain. *Journal of the Royal Statistical Society, Series A*, (forthcoming).

Hancock, R. M. and Pudney, S. E. (2013). Assessing the distributional impact of reforms to disability benefits for older people in the UK: implications of alternative measures of income and disability costs. *Ageing and Society*, 34:232–257.

Hausman, J. A., Abrevaya, J., and Scott-Morton, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87:239–269.

Idler, E. L. and Benyamini, Y. (1997). Self-rated health and mortality: a review of 27 community studies. *Journal of Health and Social Behavior*, 38:21.

Jäckle, A., Burton, J., Kaminska, O., McFall, S., and Uhrig, C. S. N. (2014). *Understanding Society. The UK Household Longitudinal Study Innovation Panel, Waves 1-6 User Manual*. ISER, University of Essex.

Kreuter, F., McCulloch, S., Presser, S., and Tourangeau, R. (2011). The effects of asking filter questions in interleafed versus grouped format. *Sociological Methods & Research*, 40(1):88–104.

Manor, O., Matthews, S., and Power, C. (2001). Self-rated health and limiting longstanding illness: inter-relationships with morbidity in early adulthood. *International Journal of Epidemiology*, 30:600–607.

Morciano, M., Hancock, R. M., and Pudney, S. E. (2014). Disability costs and equivalence scales in the older population in Great Britain. *Review of Income and Wealth*.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33:341–365.

Schroll, M., Ferry, M., Lund-Larsen, K., and Enzi, G. (1991). Assessment of health: self-perceived health, chronic diseases, use of medicine. *European Journal of Clinical Nutrition*, 45:169–182.

Sturgis, P., Thomas, R., Purdon, S., Bridgwood, A., and Dodd, T. (2001). *Comparative review and assessment of key health state measures of the general population*. Research Report. Department of Health, London.

Sweeney, K. and Furphy, M. (2008). An exercise in surveying a non-universally defined group in the population. the Northern Ireland Survey of Activity Limitation and Disability. *Journal of the Statistical and Social Inquiry Society of Ireland*, 37:217–269.

van Doorslaer, E., Wagstaff, A., and Bleichrodt, H. (1997). Income-related inequalities in health: some international comparisons. *Journal of Health Economics*, 16:93–112.

Wanless, D. (2006). *Securing Good Care for Older People: Taking a Long-term View*. King's Fund, London.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, non-linear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics*, 20:39–54.

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: The response error "correction" algorithm

Let Y_{it}^* be the (potentially misreported) binary answer of respondent i to question HEALTH in periods $t = 5, 6, 7$ and define R_{it} to be a discrete indicator taking the values 1...6 for each of the RDI responses in the order they are listed in the relevant panel of Table 6 above.[11] The free-text follow-ups to the $R_{it} = 1$ and $R_{it} = 6$ responses are varied but fall naturally into three groups, which we indicate by a discrete variable F_{it} $(t = 6, 7)$:

$F_{it} = 0$: change of mind about current at wave t
$F_{it} = 1$: definite claim of an error at wave $t - 1$
(*e.g.* "said he never had an illness", "entry error")
$F_{it} = 2$: mention of a specific medical condition that existed at $t - 1$
(*e.g.* "angina and asthma for many years", "high blood pressure 18 months")
$F_{it} = 3$: reference to intermittent nature of a long-standing condition
(*e.g.* "migraine", "asthma all my life it varies [with] the activities I do")

We construct empirical indicators of two alternative health/disability concepts: C_{it} indicates the existence of a health condition; and L_{it} indicates a health condition that limits (or would limit) normal activity

C_{it} and L_{it} are constructed using the following algorithm.

(1) Initialise $C_{it} = L_{it} = Y_{it}^*$ for $t = 5...7$, then edit C_{it}, L_{it} as follows:

(2) If exit is reported at wave 6 $(Y_{i5}^* = 1, Y_{i6}^* = 0)$:

(i) reset $C_{i6} = L_{i6} = 1$ if current report changed ($R_{i6} = 1$ and $F_{i6} = 0$)

(ii) reset $C_{i6} = 1$ if evidence of persistence ($R_{i6} \in \{2, 3, 4\}$)

(iii) reset $L_{i6} = 1$ if evidence of actual or potential limitation ($R_{i6} = 4$)

(iv) reset $C_{i5} = L_{i5} = 0$ if definite evidence of error at wave 5 ($R_{i6} = 1$ and $F_{i6} = 1$)

(3) If entry is reported at wave 6 $(Y_{i5}^* = 0, Y_{i6}^* = 1)$:

(i) reset $C_{i6} = L_{i6} = 0$ if current report changed ($R_{i6} = 1$ and $F_{i6} = 0$)

(ii) reset $C_{i5} = 1$ if evidence of pre-existence ($R_{i6} \in \{2, 3, 4\}$)

[11] Note that R_{it} can take multiple values, since the standard responses listed in Table 6 are not mutually exclusive, so the condition $R_{it} \in \{2, 3, 4\}$ or $R_{it} = 4$ appearing below is satisfied if any one of the individual's responses satisfies it.

(iii) reset $L_{i5} = 1$ if evidence that limitation could have occurred at wave 5 ($R_{i6} = 4$)

(iv) reset $C_{i5} = 1$ if definite evidence of error at wave 5 ($R_{i6} = 1$ or 6 and $F_{i6} \geq 2$

(4) If exit is reported at wave 7 ($Y^*_{i6} = 1, Y^*_{i7} = 0$):

(i) reset $C_{i7} = L_{i7} = 1$ if current report changed ($R_{i7} = 1$ and $F_{i7} = 0$)

(ii) reset $C_{i7} = 1$ if evidence of persistence ($R_{i7} \in \{2, 3, 4\}$)

(iii) reset $L_{i7} = 1$ if evidence of actual or potential limitation ($R_{i7} = 4$)

(iv) reset $C_{i6} = L_{i6} = 0$ if definite evidence of error at wave 6 ($R_{i7} = 1$ and $F_{i7} = 1$)

(5) If entry is reported at wave 7 ($Y^*_{i6} = 0, Y^*_{i7} = 1$):

(i) reset $C_{i7} = L_{i7} = 0$ if current report changed ($R_{i7} = 1$ and $F_{i7} = 0$)

(ii) reset $C_{i6} = 1$ if evidence of pre-existence ($R_{i7} \in \{2, 3, 4\}$)

(iii) reset $L_{i6} = 1$ if evidence that limitation could have occurred at wave 6 ($R_{i7} = 4$)

(iv) reset $C_{i6} = 1$ if definite evidence of error at wave 6 ($R_{i7} = 1$ or 6 and $F_{i7} \geq 2$