EEPRU

Policy Research Unit in Economic Evaluation of Health and Social Care Interventions

An empirical comparison of well-being measures used in UK

Clara Mukuria, Tessa Peasgood, Donna Rowen, John Brazier

RR0048
November 2016

Correspondence to: Clara Mukuria, ScHARR, University of Sheffield,
c.mukuria@sheffield.ac.uk

The University Of Sheffield.

THE UNIVERSITY of York

The Policy Research Unit in Economic Evaluation of Health and Care interventions is funded by the Department of Health Policy Research Programme. It is a collaboration between researchers from the University of Sheffield and the University of York.

The Department of Health's Policy Research Unit in Economic Evaluation of Health and Care Interventions is a 5 year programme of work that started in January 2011. The unit is led by Professor John Brazier (Director, University of Sheffield) and Professor Mark Sculpher (Deputy Director, University of York) with the aim of assisting policy makers in the Department of Health to improve the allocation of resources in health and social care.

EXECUTIVE SUMMARY

Background

A number of different, yet related, measures of subjective well-being (SWB) and health are used across government departments. Under its Measuring National Well-being Programme, the Office of National Statistics (ONS) has adopted the use of the short Warwick Edinburgh Mental Well-being Scale (SWEMWBS) and the General Health Questionnaire (GHQ-12), which is a mental health screening measure, as well as four summary subjective (personal) well-being questions which ask about life satisfaction, happiness and anxiety yesterday, and worthwhileness (the ONS-4). In addition to the measures used within the ONS framework, the National Institute for Health and Care Excellence (NICE) currently prefer the EQ-5D, a measure of health-related quality of life (HRQoL), in the assessment of medical technologies and public health interventions,[NICE, 2013a] while social care guidance includes measures of capability and need: the Investigating Choice Experiments Capability Measures for Older people/ Adults (ICECAP-A and ICECAP-O) and the Adult Social Care Outcomes Toolkit (ASCOT).[NICE, 2013b] There is limited evidence on how these measures relate to each other, which causes difficulty in the comparison of results across datasets and evaluations containing different measures, as well as for informing decisions across sectors. Given that these measures are used to inform policy making throughout Government, it is important to better understand how these measures compare.

The Department of Health has asked the Policy Research Unit in Economic Evaluation of Health and Care Interventions (EEPRU) to undertake a conceptual and empirical comparison of the following six commonly used measures of health and well-being: SWEMWBS, GHQ-12, ONS-4, ICECAP, ASCOT and EQ-5D. This report summarises psychometric analysis including factor analysis which sought to compare the ONS-4, the SWEMWBS/WEMWBS, the GHQ-12, the ICECAP-A or ICECAP-O, ASCOT, the EQ-5D and the SF-6D. The report also takes into consideration additional measures of SWB found within the datasets to shed further light on these comparisons and the concepts behind the measures. It addresses three related questions: 1) whether the well-being measures measure the same or different constructs related to the underlying theoretical foundations;[see Peasgood et al, 2014] 2) whether or not separate positive and negative well-being measures are required; and 3) what the potential impact of using well-being measures in the evaluation of health-care interventions would be. The questions relate to the more specific question of whether there is redundancy if both the GHQ-12 and the S-WEMWBS are measured. GHQ-12 has both negative and positive items while the S-WEMWBS focuses only on positive well-being questions. The key question is whether the negative questions provide additional, policy relevant, information to the positive well-being questions.

Method

Five datasets were used for the analysis:

- Health improvement and Patient Outcomes (HIPO): a large UK patient dataset that collected SWB and health data in inpatients recently discharged from hospital in 2014;
- Multi Instrument Comparison (MIC): a survey collected from online research panels in 2012 from six countries, including the UK;
- South Yorkshire Cohort over 65 (SYC65): a general population sample recruited from a cohort that was recruited from general practitioners in Yorkshire and Humber;
- Understanding Society (USoc): wave 1 (2009-10) and wave 4 (2013-14) of the UK household panel;
- Health Survey for England (HSE): the 2010 wave of the general population health survey.

Classic psychometric analysis assessing the relationship between these measures was undertaken. This included exploring summary statistics of the different SWB (ONS-4, GHQ-12, S-WEMWBS, ICECAP-A/O) and health (EQ-5D and SF-6D) and social care (ASCOT) measures and looking at correlations between these measures. Factor analysis was used to assess whether or not the measures covered more than one dimension. The relative ability of the SWB measures to discriminate between groups with known differences in health compared to health measures was also tested using effect sizes from regression analysis (eta squared). The groups were defined by diagnosis or self-reported health problem.

HIPO, SYC65 and MIC allowed comparisons of ONS-4 and the health measures; USoc and HSE allowed comparison of GHQ-12 and S-WEMWBS and one of the 2 health measures. ICECAP-A was only available in MIC, while ICECAP-O was available in SYC65 which also had ASCOT and the WEMWBS. None of the datasets covered all the measures and ONS-4 and ICECAP measures could not be compared to GHQ-12, but there were a number of other single item SWB questions included in the analysis of the datasets. HIPO was a patient dataset whereas SYC65, MIC, USoc and HSE were general population datasets with self-reported conditions.

Findings

Convergence of SWB measures

The different SWB measures were moderately to strongly correlated with each other. The positive SWB items and measures (ONS-4 life satisfaction, happiness, worthwhile, ICECAP-A/O) were strongly correlated, with a slightly weaker relationship between positive and negative items. The distribution of scores for negative items showed fewer respondents reporting problems in the negative SWB items

compared to positive SWB items. This applied to the ONS-4 anxiety question, the six GHQ negative items and other negative SWB items available in the datasets used. The differences in distribution suggest that the negative items are not the mirror opposite to positive items. Lower correlations, differences in effect size and potential alignment to different latent factors may be influenced by these distribution differences.

Exploratory factor analysis (EFA) commonly identified positive and negative factors, and factors that lined up with the different instruments regardless of whether items were tapping into the same concepts. For example, factor analysis of the GHQ-12 and S-WEMWBS resulted in two positive SWB factors linking to positive items from the GHQ and S-WEMWBS items respectively, and one factor linking to the negative GHQ items. The grouping of items into factors which reflect the different instruments rather than the underlying constructs does not generate confidence in the method. More sophisticated confirmatory factor analysis (CFA) in which data are modelled as containing underlying SWB factors with additional methods factors gave better model fit across all five datasets. This suggests that differences between positive and negatively worded items instruments are a result of differences in measurement responses to positively and negatively worded items rather than differences in constructs.

However, there were differences in the way negative SWB questions for the GHQ were associated with unemployment, with larger effect sizes for the negative SWB questions than for the positive. There are a number of possible interpretations for this:

1. Positive and negative emotions are to some degree independent, tapping into separate dimensions
2. The measurement error is different for positive and negative items
3. Positive and negative items focus on different parts of the scale of a single construct, and relationships with other characteristics vary across the scale (i.e. unemployment may be very damaging to someone with already poor SWB, but have minimal impact on those with high SWB).

It is difficult to confidently distinguish between these possibilities.

Based on these results it is difficult to judge whether or not the ONS should include GHQ-12 alongside the S-WEMWBS. Their performance and acceptability in groups with particularly low SWB and mental health problems would be an additional criteria to consider and an area requiring future research.

Comparing SWB and health measures

Generally, effect sizes for physical health conditions were much smaller for SWB measures than for the EQ-5D and the SF-6D. Results were mixed for depression or mental health: GHQ-12 (and GHQ negative) had a larger effect size than the health measures; ICECAP-O/A had a larger effect size than EQ-5D-5L; S-WEMWBS had a larger effect size than EQ-5D-3L and about the same as the SF-6D. However, the single ONS-4 items generally had lower effect sizes for depression and mental health than the health measures.

This finding was confirmed when reported disability was used, with SWB measures showing greater sensitivity to memory or concentration problems than the health measures while the reverse was true for a physical health disability. Panel data supported the finding that GHQ and WEMWBS were better at discriminating between groups with depression or limitations in memory/concentration than the SF-6D while they had lower discriminative power for physical health conditions or limitations. As physical health is only one aspect of SWB it is in line with expectations that they would show smaller effect sizes for the health conditions, with the exception of depression. If SWB measures were to be used to evaluate health care interventions they may show greater sensitivity to changes in mental health, but this depends upon the measure and would need confirmation in panel data. This data suggests that single item SWB measures such as the ONS-4 would still be less sensitive to changes in mental health than existing health measures; however, the relative importance of physical to mental aspects of health would change.

Limitations

There are a number of limitations. None of the datasets contained all the measures together. ONS-4 was compared with a number of SWB and health measures but not with the GHQ-12. An assessment of the S-WEMWBS against negatively worded SWB items was limited to the GHQ-12 which has ambiguous responses, for example it is difficult to know what respondents mean by 'no more than usual'. Only the USoc and SYC65 data included change over time, with only two time points. Differences in the mode of administration and wording of questions including classification of health conditions also limit comparisons across the datasets.

Conclusions and Recommendations

The findings of this report provide evidence on the relationship between different SWB measures and commonly used health measures. Although there are differences in some of the SWB measures, including positive and negative measures, it is not possible to recommend at this stage whether or not

one or more measures should be excluded from the group in current use. If the aim is to provide a measure of SWB that can be compared across individuals, one possibility would be to replace the GHQ-12 because the response options are ambiguous. However, S-WEMWBS may not be an appropriate replacement because it does not contain negative items, while the ONS-4 rely on single item measures which may not perform as well as aggregate measures.

The implications of any move towards using SWB to evaluate health policy needs to be carefully considered. SWB measures, including those focusing on psychological well-being, are far less sensitive to physical health conditions. Therefore, moving to SWB would result in a substantial increase in the weight given to mental health compared with physical health conditions. This would also have dramatic implications for sample sizes required to detect changes in health.

TABLE OF CONTENTS

1. BACKGROUND

In recent years there has been increasing interest in the measurement of well-being, for example in the UK's Office of National Statistics (ONS) cross-government Measuring National Well-being Programme. A number of different, yet related, measures of well-being and health are used across government departments. This includes four summary subjective (personal) well-being questions which ask about life satisfaction, happiness yesterday, anxiety yesterday and worthwhileness adopted by the ONS under its Measuring National Well-being Programme (referred to here as the ONS-4). They have also adopted the use of the Short Warwick Edinburgh Mental Well-being Scale (S-WEMWBS), a measure of positive mental health, and the General Health Questionnaire (GHQ-12), a mental health screening measure. In addition to the measures used within the ONS framework, the National Institute for Health and Care Excellence (NICE) currently prefer the EQ-5D, a measure of health-related quality of life (HRQoL), in the assessment of medical technologies and public health interventions,[NICE 2013a] Social care guidance from NICE includes measures of capability and need: the Investigating Choice Experiments Capability Measures for Older people/ Adults (ICECAP-O and ICECAP-A) and the Adult Social Care Outcomes Toolkit (ASCOT).[NICE, 2013b]

Little is known about how these different health and well-being measures compare to each other. Most published studies compare a subset of the measures. A cross-sectional comparison (n=8022) of EQ-5D-5L, SF-6D (an alternative HRQoL measure) and other preference-based measures to the ONS-4 (total score of the 4 items) found that EQ-5D has less association with ONS-4 than SF-6D and accounted for less of the impact of health conditions (depression, diabetes, cancer, heart disease, arthritis, asthma, hearing loss) on the ONS-4 than the SF-6D.[Richardson et al 2015] Mitchell et al [2015] undertook a cross-sectional comparison of the EQ-5D-5L and ICECAP-A using the same dataset as Richardson et al [2015] and they found that depression had a bigger impact on ICECAP-A than EQ-5D-5L while for the other conditions, the relative impact of the condition varied depending on which measure was used. Similarly, in a study of frail older adults (n=190) EQ-5D-3L was more associated with physical limitations than broader well-being concepts than ICECAP-O and ASCOT.[van Leeuwen et al 2015] This finding is supported by other studies in older populations where EQ-5D-3L reflects impact of physical health more [Couzner et al 2013] but ICECAP-O is able to reflect outcomes beyond physical limitations such as cognition.[Davis et al 2012] However, in a longitudinal study of older adults receiving instructor-led physiotherapy (n=357), there was evidence that EQ-5D-3L reflected improvements in terms of anxiety and depression better than ICECAP-A while the reverse was true for worse health in these symptoms. [Keeley et al 2015]

There is limited evidence on how the GHQ-12, ONS-4, S-WEMWBS, EQ-5D, ASCOT and ICECAP-O/A relate to each other, which causes difficulty in the comparison of results across datasets and evaluations containing different measures, as well as for informing decisions across sectors. Given that these measures are used to inform policy making throughout Government, it is important to better understand how these measures

compare. For example, it is unclear whether the GHQ-12 is redundant if the S-WEMWBS is included alongside the ONS-4 within the ONS well-being indicator set.

The Department of Health asked the Policy Research Unit in Economic Evaluation of Health and Care Interventions (EEPRU) to undertake a conceptual and empirical comparison of these six commonly used measures of health and well-being: S-WEMWBS, GHQ-12, ONS-4, ICECAP, ASCOT and EQ-5D. This report addresses the second task by providing an empirical comparison of these measures using large datasets. It addresses three related questions: 1) whether the well-being measures measure the same or different constructs related to the underlying theoretical foundations;[see Peasgood et al, 2014] 2) whether or not separate positive and negative well-being measures are required; and 3) what the potential impact of using well-being measures in the evaluation of health-care interventions would be. The questions relate to the more specific question of whether there is redundancy if both the GHQ-12 and the S-WEMWBS are measured. GHQ-12 has both negative and positive items while the WEMWBS focuses only on positive well-being questions. A key question is whether the negative questions provide additional, policy relevant, information to the positive well-being questions. For example, it may be that they have different dimensions/constructs and so knowing only about positive well-being is not sufficient to assess well-being. Alternatively, it could be that they measure the same construct but at different points of the scale, for example negative items allow a more accurate assessment of people with low SWB. Finally, it may be that they are measuring the same thing and so only one is required. It is also useful to assess whether the recommended ONS-4 questions and other measures such as the ICECAP-A provide similar or additional information when assessing well-being.

2. METHODS

Four measures which are recommended for use as well-being or mental health measures were assessed: the ONS-4, GHQ-12, WEMWBS (and its short form the S-WEMWBS) and ICECAP-O/A, alongside a number of single item well-being questions that were in the available datasets either as individual questions or as questions within other measures. For simplicity, all these measures are referred to as subjective well-being (SWB) measures. These SWB measures were compared to two HRQoL measures, the EQ-5D and the SF-6D, both of which are generic preference-based measures used in health-care assessment. ASCOT was also included as a measure of social care related quality of life (SC-RQoL). The data and measures as well as methods of analysis are described further in this section.

2.1. Data

The analysis used five datasets which are described below.

2.1.1. Health improvement and Patient Outcomes (HIPO)

The Health improvement and Patient Outcomes (HIPO) is a large UK patient dataset. HIPO was designed to collect health and SWB data from inpatients recently discharged from hospital. Data were collected using a

prospective survey conducted in 2013-14 of inpatients at Cardiff and Vale NHS Hospitals Trust, which is a large University hospital in South Wales, UK. The survey covered most specialties, though patients with a primary mental health diagnosis are not included. HIPO surveys were sent 6 weeks after discharge, and included all subjects aged 18 years or older. The survey was linked to existing routine hospital data to provide a dataset with socio-demographic (age, gender), HRQoL (EQ-5D-5L, SF-12), well-being (ONS-4, single positive and negative SWB items, SWB-VAS) and diagnosis data. The International Classification of Diseases 10th version (ICD-10) was used to record clinical diagnosis in the hospital.[WHO, 2010] ICD is the standard method for classifying diseases and other health problems. Routine data on surgical procedures was also linked to survey data.

Ethical approval for the study was obtained from the NRES Committee North West – Cheshire.[REC REF: 12/NW/0535] 25,919 questionnaires were sent to discharged patients between September 2013 and January 2014 and 6,351 returned completed questionnaires, a response rate of 25%. Of these, 630 had more than one visit to the hospital during this period but they only completed a single questionnaire. ICD-10 code assigned to the patient was taken from the longest stay. 1,007 (16%) had missing data in one or more of the questions asked, and they were older (65.2 vs. 59.2, t_{6349}=-11, p<0.001), less healthy (EQ-5D-5L: 0.597 vs. 0.685, t_{6141} = 8.2, p<0.001) and had lower SWB (life satisfaction: 6.17 vs. 6.67, t_{6243} = 5.5, p<0.001; anxious: 6.92 vs. 7.51, t_{6246} = 5.6, p<0.001) than those with no missing data. Four ICD condition groups (Chapter 5 - mental health disorders, Chapter 8 - ear and mastoid diseases, Chapter 15 - pregnancy, and Chapter 17 - congenital malformations) had small samples (n<30) and these were merged into a miscellaneous ICD group for the analysis (Appendix Table 1). Mean (SD) age was 59.2 (16.39) and 50% were female for those without missing data.

2.1.2. Multi Instrument Comparison (MIC)

The Multi Instrument Comparison (MIC) dataset is a cross national survey which collected data online in 2012 from six countries: Australia, Canada, Norway, Germany, UK and USA. Respondents were members of panels that had agreed to participate in online research. Respondents answered a main questionnaire consisting of: SWB (personal well-being index (PWI), ONS-4, satisfaction with life survey (SWLS)), generic preference-based measures of HRQoL (EQ-5D-5L, AQoL-8D and AQoL-4D, HUI3, 15D, QWB-SA), a non-preference based measure of HRQoL (SF-36), self-complete time trade off (TTO) of their own health, and ICECAP-A as well as demographic questions. In addition, respondents self-reported whether they had depression, hearing loss, asthma, COPD, diabetes, arthritis, heart disease, stroke or cancer and completed condition specific measures. The generic heath measures were presented in a random order to different respondents. The study aimed to recruit 9,150 respondents (healthy: 2,100; disease sample: 7,050). Respondents were invited to take part from an online panel with screening questions used to identify the condition group for the disease groups until a specified quota within each country was reached (150 per disease group). Those who did not have a condition received a

visual analogue scale (VAS) health question which was used to filter out individuals who had a VAS score of 70 (initially 65) or less. Respondents were required to complete all the included measures. Planned edit procedures were used to ensure that responses were consistent. The final sample size was 8,022 (healthy public: 1,760; disease sample: 6,262) and included only respondents who completed the whole survey as those who did not complete the survey were excluded. 14 respondents were excluded as their HRQoL measures could not be linked to SWB measures. One individual did not have SF-36 data and the stroke sample was small (n=23) so these respondents were excluded from the analysis. Norwegians (n=1,176) did not complete all the measures and they were also different from the rest of the sample (life satisfaction: 6.83 vs. 6.03, t_{7984} = 9.4, p<0.001; anxious: 7.58 vs. 6.35, t_{7984} = 13.5, p<0.001; EQ-5D-5L: 0.794 vs. 0.728, t_{7984} = 9.2, p<0.001), therefore they were excluded from the analysis leaving a sample size of 6,808. Mean (SD) age was 51.2 (15.11), 54% were female and 21.6% were in the healthy group (Appendix Table 1).

2.1.3. South Yorkshire Cohort Over 65 (SYC65)

The South Yorkshire Cohort over 65 (SYC65) is a survey that is undertaken with respondents aged 65 or over recruited from the South Yorkshire Cohort (SYC, now Yorkshire Study), a large existing general population cohort. The SYC uses the cohort multiple randomised controlled trial design which allows other studies to recruit from it by targeting respondents with specific characteristics who have self-identified as happy to participate in future research.[Relton et al 2011] The initial cohort was recruited between 2010 and 2012, from South Yorkshire, United kingdom (Barnsley, Rotherham and Doncaster). All patients registered in GP practices aged 16 to 85 were recruited. 13,760 participants in this initial cohort have agreed to take part in future studies, 6,600 of whom are individuals aged 65 and over years with at least one self-reported long term condition (LTC). Between November and December 2014, 3,575 SYC members were invited to participate in a survey that covered health and wellbeing based on age, gender, past health status and different LTCs. 1,749 responded giving a response rate of 48.9%. Participants were asked questions regarding their health (EQ-5D-5L), well-being (ONS-4, WEMWBS, ICECAP-O) and social care (ASCOT) as well as questions related to long-term conditions and health and social care service use. At follow-up they were asked the same health and wellbeing questions as well as whether they had a new condition. Ethics review was provided by the School of Health and Related Research in the University of Sheffield.

The overall sample size was 1,749, but this included a number of respondents with missing data. Respondents with complete responses for the well-being and HRQoL measures are used here, totalling 1,593 respondents. Those with missing data were older (75.3 vs. 72.6, t_{1747}=-5.4, p<0.001), less healthy (EQ-5D-5L: 0.672 vs. 0.728, t_{1727} = 3.0, p<0.05) and had lower SWB (life satisfaction: 7.30 vs. 7.67, t_{1742} = 2.38, p<0.05; anxious: 6.62 vs. 7.30, t_{1740} = 3.01, p<0.05) than those with no missing data. Mean (SD) age was 72.6 (5.77), 51% were female and

27% did not report any of the conditions listed for those without missing data (Appendix Table 1). Only 102 (9%) and 12 (1%) reported a new physical or mental health condition at follow-up, respectively.

2.1.4. Understanding Society (USoc)

Understanding Society (USoc) is a UK panel which annually surveys the same representative panel of households. USoc started interviews for wave 1 in 2009-2010 and replaced a previous 18 year-long panel survey, the British Household Panel Survey (BHPS) which ran from 1991. Wave 1 and wave 4 of USoc was used in the analysis as these are the only waves with multiple SWB measures included. Respondents in wave 1 and wave 4 were asked a large number of questions relating to their socio-demographics and SWB including GHQ-12, S-WEMWBS, a single item question on life satisfaction ("How dissatisfied or satisfied are you with your life overall?"), and the SF-12[1] which allows the generation of the SF-6D.

In both wave 1 and wave 4 respondents were asked whether they had a long-standing illness or disability, and whether this resulted in substantial difficulties with any of the following areas of their life: mobility (moving around at home and walking), carrying or moving objects, manual dexterity (using hands to carry out everyday tasks), communication or speech problems, memory or ability to concentrate, learn or understand, recognising physical danger, physical co-ordination (e.g. balance), difficulties with own personal care, continence (bladder and bowel control), hearing (apart from using a standard hearing aid), sight (apart from wearing standard glasses), other.

Respondents were asked in wave 1 whether a doctor or other health professional had ever told them that they have any of 17 health conditions and whether they still have the condition (asthma, arthritis, congestive heart failure, coronary heart disease, angina, heart attack or myocardial infraction, stroke, emphysema, hyperthyroidism or an over-active thyroid, hypothyroidism or an under-active thyroid, chronic bronchitis, any kind of liver condition, cancer or malignancy, diabetes, epilepsy, high blood pressure, clinical depression). In subsequent waves respondents are asked whether they have been newly diagnosed with any of the health conditions since the previous interview date. However, they were not asked whether they were still experiencing conditions reported in prior waves. To simplify the reporting of the analysis and ensure reasonable group sample sizes some similar conditions were combined.

The overall sample size in wave 1 was 50,994 but this included a large number of missing data responses. Respondents with complete responses for the SWB and HRQoL measures are used here, totalling 37,602 respondents. Those with missing data were older (48.2 vs. 45.3, t_{47730}=14.2, p<0.001), less healthy (SF-6D (SF-12): 0.763 vs. 0.796, t_{47482} = -20.2, p<0.001) and had slightly lower SWB (GHQ-12: 11.57 vs. 11.02, t_{39698} = 4.58,

[1] In wave 1 the SF-12 is asked during the CAPI (Computer Assisted Personal Interviewing) interview, in subsequent waves it is asked within the self-completion section. This makes a difference to the number of missing values.

p<0.001) than those with no missing data. Mean (SD) age in wave 1 was 45.3 (17.75), 56% were female and 53% did not report any of the conditions listed for those without missing data (Appendix Table 2).

2.1.5. Health survey for England (HSE)

The Health Survey for England (HSE) is an annual general population survey conducted in England via interview since 1991 that examines the nation's health. The survey is not a panel, which means that it does not survey the same respondents each year and link their responses. Participants are asked a large number of questions about their health, and the data used here are: GHQ-12, WEMWBS, a single item on happiness ("Taking all things together, on a scale of 0 to 10, how happy would you say you are?"), EQ-5D-3L and EQ-VAS. Respondents were asked whether they had a longstanding illness and the type of longstanding illness. Illnesses were classified into ICD-10 chapters. The data used here was collected between January and December 2010, with a household response rate of 66%. The overall sample size was 14,112, but this included a large number of missing data responses and proxy respondents. Respondents with complete responses for the well-being and HRQoL measures are used here, totalling 5,709 respondents. Those with missing data were older (56.0 vs. 47.7, t_{7669}=16.6, p<0.001), less healthy (EQ-5D-3L: 0.802 vs. 0.860, t_{7330} = -8.75, p<0.001) but had only slightly lower SWB (GHQ-12: 11.16 vs. 10.76, t_{7470} = 2.94, p<0.05) than those with no missing data. Mean (SD) age was 47.7 (17.83), 56% were female and 58% did not report any of the conditions listed (Appendix Table 2).

2.2. Measures – subjective well-being

2.2.1. ONS-4

The ONS-4 questions are: "Overall, how satisfied are you with your life nowadays?" (0 not at all to 10 completely), "Overall, to what extent do you feel the things you do in your life are worthwhile?", "Overall, how happy are/were you today/yesterday?" and negative affect "Overall, how anxious are/were you feeling today/yesterday?" These measures of SWB have been recommended for use in the UK.[Dolan et al, 2012] In the analyses reported here, the ONS anxious question was recoded so that 0 was "completely anxious" and 10 was "not at all" in order for higher values to represent higher SWB in a similar way to the other three ONS SWB questions. In addition to looking at these data as four single items, an aggregate score was constructed in which each response was equally weighted. Initial analysis indicated that the 'anxious' question was different form the other three so a second aggregate score was created excluding this question. Aggregation was done purely for the purpose of exploring the data and is not an approach used by the ONS.

2.2.2. GHQ-12

The GHQ was developed as a first-stage screening tool to measure mild somatic and psychological symptoms in a non-clinical environment and to identify those in need of psychiatric care.[Goldberg and Williams, 1988] The questionnaire focuses on two major areas: the "inability to carry out one's normal 'healthy' functions, and the appearance of new phenomena of a distressing nature",[Goldberg and Hillier, 1979: p139] the aim being to identify individuals who are disturbed or altered from their usual self. The original scale comprised of 60 items, but 30, 28, 20 and 12-item versions have since been developed. The GHQ-12 includes six positive and six negative questions and a choice of four response options for each in which the presence or intensity of the state over the last few weeks is related to its usual frequency or intensity. Negative items have response options of "not at all / no more than usual / rather more than usual / much more than usual", and positive items have response options of "more so than usual / same as usual / less so than usual / much less than usual". Scoring can adopt a number of different forms, the three most common scoring methods being: the GHQ 'caseness' score (scored as 0-0-1-1) for positive questions and negative questions representing the number of the 12 symptoms present; scoring each item on a four point scale (0-1-2-3) to give a Likert score out of 36; or a 'corrected' binary score (CGHQ) which, unlike the caseness score, takes 'same as usual' for the negative items as an indication of the presence of a symptoms.[Goodchild and Duncan-Jones, 1985] It is the Likert scoring that is reported here, with higher scores indicating poorer mental health. The caseness scoring was also analysed but there were no differences in the results so these are not reported (results available from the authors).

We also scored the six positive and six negative items separately using Likert scoring purely for data exploration purposes (referred to as GHQ positive and GHQ negative, respectively). Although the GHQ-12 was developed to assess mental health, it has been used to assess SWB in the literature as the questions it asks refer to hedonic and flourishing aspects of well-being.[Peasgood et al, 2014] Therefore, it is included to assess whether it measures separate constructs to the other measures of SWB.

2.2.3. WEMWBS and S-WEMWBS

The WEMWBS was developed from the Affectometer by the Universities of Warwick and Edinburgh.[Kammann and Flett, 1983] The scale aimed to be able to identify levels of positive mental health in the general population and drew from a number of different conceptions of well-being, including hedonic (feelings) and flourishing accounts (psychological functioning and self-realisation).[Tennant et al, 2007] The full version asks for time spent in 14 positive states over the last two weeks with five response categories ranging from 'all of the time' to 'none of the time'. Responses are totalled giving a minimum score of 14 and a maximum of 70, with a higher score indicating higher mental health. A shortened 7-item version, the S-WEMWBS, has also been derived using Rasch and has items on optimism, usefulness, feeling relaxed, thinking clearly, dealing with problems,

feeling close to others, and being able to make up one's own mind.[Stewart-Brown et al, 2009] The S-WEMWBS can be scored by linking to the Rasch scale [Stewart-Brown et al, 2009] but we use the simple aggregate scoring here in line with the full version, giving a possible range from 7 to 35.

2.2.4. Investigating Choice Experiments Capability Measure (ICECAP-O and ICECAP-A)

The ICECAP-O and ICECAP-A (Investigating Choice Experiments Capability Measure for Older people/Adults) are capability measures that draw upon Sen's capability theory. The ICECAP-O was developed for use in older populations,[Grewal et al, 2006] while the ICECAP-A was developed for use in adults.[Al-Janabi et al, 2012] The ICECAP-O has 5 items that can take one of four levels: attachment (can have the love and friendship: all, a lot, a little, not any), security (can think about future without concern: any, a little, some, a lot), role (able to do the things that make me valued: all, many, few, unable), enjoyment (can have enjoyment and pleasure that I want: all, a lot, a little, cannot) and control (able to be independent: completely, in many things, few things, unable). The ICECAP-A has five items that can take one of four levels: stability (able to feel settled and secure: in all areas of life; many areas; few areas; unable to), attachment (can have love, friendship and support: a lot; quite a lot; a little; unable to), autonomy (able to be independent: completely; in many things; in few things; unable to), achievement (able to achieve and progress: in all aspects of life; in many aspects; in few aspects; unable to) and enjoyment (able to have enjoyment and pleasure: a lot; quite a lot; a little; unable to).[Al-Janabi et al, 2012] Both the ICECAP-A and ICECAP-O can be scored based on a best-worst scaling (BWS) exercise with the general public [Flynn et al, 2013]. The values have not been anchored on the QALY scale, where zero is equivalent to being dead, but on a 0-1 scale where zero is the worst (i.e. no capability on any dimension) and one the best (full capability on all dimensions) state defined by the measure. [Coast et al, 2008; Flynn et al, 2013]

2.2.5. Single-item subjective well-being measures

HIPO and SYC65 had positive SWB worded questions on feeling content, ability to do things, looking forward to tomorrow, having supportive relationships, contributing to others' happiness, doing enjoyable things and life going well; most of these items were scored from strongly disagree (0) to strongly agree (10), apart from the feeling content question which was scored like the ONS-4. A SWB-VAS which had the same layout as the EQ-VAS was also included as a single item measure of SWB, where instead of HRQoL it focuses on life overall, i.e. 'We would like to know how good or bad your life is' with response options - '100 means the best life you can imagine' and '0 means the worst life you can imagine'.

HIPO also had negative items on feeling tired, lonely, angry, and bored, with scoring similar to the ONS-4 anxious item. Negative items were recoded so that at 0 was "completely" and 10 was "not at all" (i.e. higher is

always better). Aggregate measures of all positive SWB items and all negative SWB items (excluding those from the ONS-4) were created by summing across items, hence weighting each item and response category equally.

MIC data included the AQoL-8D which is a preference-based measure that includes happiness and mental health dimensions.[Richardson et al, 2013] Six positive (happiness, enthusiasm, pleasure, enjoying close relationships, feeling control and contentment) and six negative (despair, worry, depression, feeling isolated, anger and sadness) items were drawn from the AQoL. Summary scores were created for the positive and negative well-being items, separately, by summing across the items, as done for the HIPO.

The Health Survey for England (HSE) contains a single question on overall happiness. The wording for each of these items can be found in Appendix 2.

2.3. Measures – HRQoL and SCRQoL

2.3.1. EQ-5D

The EQ-5D is a generic preference-based measure of HRQoL which is preferred by NICE in economic evaluation and was therefore used as a comparator to the SWB measures included in this analysis. The EQ-5D consists of a health state classification system with five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. There are two versions: the 3 level version which has three severity levels for each dimension and the more recently developed 5 level version. The 3-level version has utility values elicited by TTO from the general population that range from -0.594 to 1,[Dolan, 1997] and a cross-walk algorithm can be used to generate EQ-5D-3L utility values using the same valuation survey for the EQ-5D-5L.[Herdman et al, 2011; van Hout et al, 2012] EQ-5D also has a visual analogue scale (VAS) which asks respondents "…how good or bad" their health is on a scale from 0 - worst health to 100 - best health that they can imagine. EQ-5D-3L was used in USoc and HSE while EQ-5D-5L was used in HIPO, MIC and SYC65.

2.3.2. SF-6D

SF-6D is a health state classification system derived from the Short Form 36 or 12 (SF-36/SF-12) which are widely used generic non-preference-based measures of HRQoL.[Brazier et al, 1998; Brazier et al, 2002; Brazier and Roberts, 2004] The SF-6D has six dimensions: physical functioning, role limitations, social functioning, pain, mental health and vitality, with 3 to 6 levels of severity depending on the dimension and whether it has been derived from the SF-36 or SF-12. It has utility values from UK general population using standard gamble that range from 0.301/0.345 to 1 (SF-6D derived from SF-36 and SF-12 respectively). It was included in the analysis to provide additional information from a health perspective as it covers additional health domains that are not included explicitly in EQ-5D, namely, social functioning and vitality. HIPO and USoc includes the SF-6D based on the SF-12, while MIC includes the full SF-36 which could be used to derive both versions of SF-6D referred to as SF-6D (SF-12) and SF-6D (SF-36).

2.3.3. Adult Social Care Outcomes Toolkit (ASCOT)

The Adult Social Care Outcomes Toolkit (ASCOT) is a measure of SC-RQoL that is designed to assess the extent to which an individual's social-care needs and wants are being met.[Netten et al, 2012] ASCOT has eight dimensions: five reflecting basic social-care related needs (accommodation cleanliness/comfort, safety, food and drink, personal care, being treated with dignity), and three reflecting higher order concerns (control over daily life, social participation, and involvement/occupation). The development of ASCOT was also influenced by Sen's theory of capabilities], which proposes that it not just an individual's attainments and functioning that are of importance in policy decisions but also the choices that were available to them. Whilst evidence from social care users implied that opportunities or capabilities were valued, *actual* functioning was also seen as important. The notion of capabilities was therefore incorporated by using a top response level for each item that asks whether the individual is in their perceived 'ideal' state for the domain. This is followed by three levels distinguishing the degree to which respondents' actual needs are met (for example 'My home is: as clean and comfortable as I want, is adequately clean and comfortable, not quite clean or comfortable enough, or not at all clean or comfortable) (see Appendix 2). This implies that whilst capabilities may matter, an individual with low levels of functioning can be judged by society as having an unacceptable level of need regardless of whether they personally recognize this to be the case. [Netten et al, 2012] The dignity dimension relates specifically to those receiving care and is set at 'no unmet needs' for those who do not receive any care. There are two methods of scoring the instrument. One is to use scores developed from a general population survey using best-worst scaling (BWS); another is to anchor these BWS scores onto the QALY scale, where zero is for states equivalent to being dead, using TTO values for a sample of states. The second approach was used to score the ASCOT here.[Netten et al, 2012]

2.4. Analysis

The aim of the analysis was to understand the similarities and the differences between the SWB and HRQoL measures. The key questions were: whether they measured empirically different constructs given their different theoretical bases; whether negative items were required alongside positive items; and what the impact of using SWB instead of health in health care assessment would be. To answer these questions, psychometric analysis including factor analysis was undertaken, as described further in this section. To enable meaningful comparisons within datasets, only respondents with complete data across the SWB and HRQoL measures in that dataset are included in the analyses.

2.4.1. Summary statistics and acceptability

Summary statistics (mean, standard deviation (SD), median, maximum and minimum) were used to provide an overview of the measures in each dataset. Differences in summary statistics of measures on similar scales, for

example the ONS questions on life satisfaction and happiness, may indicate that they are measuring different concepts or that they may be measuring the same concept but focusing on different parts of the scale (such as high SWB or low SWB). Distribution of the scores across SWB and HRQoL was assessed using histograms to help better understand the differences between measures.

The level of missing data was also assessed. A high level of missing data compared to levels of missing data in other completed measures within each dataset provides a rough indication that a measure may not be acceptable. However, comparison between datasets is difficult due to the different administration modes and content of questionnaires which may affect the level of missing data.

Floor and ceiling effects, the proportion of respondents at the worst (lowest SWB/HRQoL) and best (highest SWB/HRQoL) score in each measure, also provides a measure of potential similarities. High floor/ceiling effects[2] mean that measures may be insensitive to deteriorations/improvements over time or differences between groups. In general population samples such as USoc and HSE, higher levels of ceiling effects may be expected, so the focus is on the relative comparison of measures within the datasets not just the absolute values.

2.4.2. Validity

Validity analysis assesses the extent to which an instrument measures what it is intended to measure by comparing them to a 'gold standard' or using appropriate indicators.[Streiner and Norman, 2008] In the absence of a 'gold standard' SWB measure, comparisons were made between measures as well as the extent to which measures were able to reflect differences in SWB and health using appropriate indicators. The aim was to identify any overlaps or differences between the measures in order to assess whether they were measuring the same thing.

2.4.2.1. Convergent validity (Correlations between measures)

Convergent validity assesses the strength of the relationship between measures using correlation analyses.[Streiner and Norman, 2008] Spearman correlations were undertaken for all the SWB and health measures. Strong correlations indicate that the instruments measure related factors. Correlations are considered weak if scores are <0.3, moderate if scores are ≥0.3 and <0.5, and strong if scores are ≥0.5.[Cohen, 1992]

These correlations need to be interpreted with caution. They may suffer from either shared method variance or random measurement error. Shared method variance means a high correlation may be attributed to the method of measurement rather than the underlying construct.[Podsakoff et al, 2003] It may arise from similar

[2] Note that for the GHQ-12, high scores indicate poor SWB. GHQ-12 scores were not reversed as the measure is commonly used in this way.

self-report biases (such as social desirability, avoiding the ends of scales, or the impact of external factors at the time of questionnaire completion), similar recall biases, cultural biases, and the impact of current mood. Method variance will inflate the correlations between our measures. This effect may be exacerbated where response scales are very similar. On the other hand, random measurement error will result in a deflation of correlation between measures because differences between scores will be due to the error as well as actual differences in what is being measured. Random measurement error is reduced when a construct is measured using more than one question or item.

2.4.2.2. *Known group validity (Effect sizes)*

A key factor in the performance of these measures is their ability to distinguish between groups in which there are known differences. One option is to look at differences in the means of the various well-being measures between two groups of people (one say with a health condition (x) and one without) while addressing the distribution of different SWB measures. Difference based effect sizes (such as Cohen's d) allow standardised comparisons of means to be made which take into account the distributions.

However, effect sizes such as Cohen's d do not address the fact that the group with the health condition (x) may have other characteristics that differ to the group without the condition, for example they are likely to be older. SWB measures have a strong association with age, typically showing a U-shaped relationship [Deaton, 2007] which means differences in SWB for groups with a health condition may be biased by age.

An easy way to control for important covariates (such as age and gender) is to use regression analysis. Rather than being based on mean differences, regression analysis generates effect sizes based on the amount of 'explained' variance, the most common being Eta^2 which is used here. Eta^2 is the proportion of the total variation in the dependent variable that can be attributed to an independent variable (ranging from 0 to 1)[3]. The interest here is in making comparisons between the effect sizes (Eta^2) within the same dataset. Ordinary least squares regressions were undertaken with Eta^2 effect sizes calculated for all the cross-sectional analysis. Separate regressions were undertaken for each SWB measure of interest with a number of independent variables where there were expected to be differences. This included health conditions compared to those with no condition and unemployed compared to everyone else (not just those who were employed). Socio-demographic differences were also taken into account, including: age, age squared, gender and whether respondents were married or not. Age squared was included to allow for the U-shaped relationship between age and SWB.

[3] Eta^2 = Sum of Squares of the effect / Sum of Squares of the total. Eta squared is a measure of strength of association. In small samples, it is upwardly biased. However, it is used in this study as large samples are used in each dataset.

In order to demonstrate the magnitude of the effect sizes in a more familiar metric, Cohen's d effect sizes[4] are additionally reported for self-reported health conditions in the USoc data (see Figure 1 below). These are based on matched data (one-to-one matched using three age categories and gender). For the Cohen's d effect size 0.2 is usually taken to be a small effect, 0.5 considered medium and 0.8 large. [Cohen, 1992]

Analysis based on regressions on cross-section data may suffer from bias arising from unobserved individual effects. Individuals with particular personality traits or a particular style of responding to questionnaires may have unobserved effects which correlate both with our outcomes of interest and covariates (such as self-report of depression). In USoc we can explore whether this bias is impacting upon our findings. Firstly, we look at change in the outcome measures based on differences between wave 1 and wave 4 and whether one of the health conditions was newly diagnosed any time between wave 1 and wave 4. An OLS regression was used with change in the outcome measure as the dependent variable (e.g. wave 4 SWB score – wave 1 SWB score), and effect sizes reported as Eta^2. We are not able to use a first-difference or fixed effects model here as we do not know whether a previously reported condition is still experienced (some could be assumed to be chronic such as emphysema, others, such as depression are more episodic), hence we rely only upon relative differences in effect sizes for any newly diagnosed conditions (hence the effect size is limited here to deteriorating health).

The USoc data on self-reported disabilities is asked in a similar manner in wave 1 and wave 4, hence we know whether individuals experience or do not experience the disability at the two time periods. We are therefore able to use a fixed effects regression which controls for any unobserved time-invariant individual effect. We report the Cohen's f^2 effect sizes which is an appropriate effect size for a fixed effects regression[5]. We use the models based on panel data as robustness checks to consider whether the results drawn from cross-section differ once we address the unobserved individual effect.

2.4.3. Factor analysis

Factor analysis aims to identify underlying unobservable (latent) variables or domains, that is, factors that cause the observed variables (or item responses) to vary together. Each questionnaire item is presented as a linear combination of the identified factors plus an error. The factor 'loadings' (these are effectively the parameters of the linear function and are between 0 and 1) show which factor (or factors) each item is correlated with. We began with a simple exploratory factor analysis (EFA), and drawing on the results from

[4] Cohen's d = (Mean outcome of group with health condition – Mean outcome of group without condition)/Pooled standard deviation. In this case all individuals not reporting the condition of interest were used as the non-condition group.

[5] Cohens f^2 has a similar interpretation to Eta^2 but allows adjustment for the variance attributed to the individual fixed effects [Selya, 2012]

the EFA and theoretical considerations we developed and tested more sophisticated confirmatory factor analysis (CFA). All analysis was conducted in MPlus.

2.4.3.1. *Exploratory factor analysis (EFA)*

EFA was used to explore the data and to find the smallest number of interpretable factors needed to explain the correlations among the set of health and well-being variables. This is exploratory in the sense that it places no structure on the linear relationships between the observed variables and the factors. The EFA was used to inform subsequent confirmatory factor analysis (CFA).

Items identifying aspects of physical health were initially included in the EFA to help understand partly whether physical health does clearly separate from mental health and well-being items, and partly to reveal other items that loaded (or cross-loaded) onto physical health. Items that are theoretically physical health were not subsequently included in the CFA. EFA was then undertaken on the mental health and well-being items as well as any items that were not clearly physical health. Items were also dropped if less than three items loaded to a clearly separate factor, with a clear conceptual interpretation, as they would not be well identified within a confirmatory model. [Costello and Osborne, 2005]

Variables on 0 to 10 scale were treated as continuous, all others were treated as categorical. Categorical variables are found within every dataset hence we used an appropriate weighted least squares robust estimator (wlsmv). Where any factor loadings arose that were above 1 we considered dropping items that were duplicated (e.g. SF-36 has an item on climbing one flight of stairs and an item on climbing several flights of stairs).

To ease interpretation of the factors we rotated the factor matrix using rotation suitable to having correlated factors (promax - oblique). To identify the number of well-being factors, a number of criteria were considered including:

- The number of factors with eigenvalues greater than one (Kaiser-Guttman criterion). However, this is only used to get a feel for the data as the cut off of one is not theoretically grounded for correlated factors.
- The number of factors appearing before the natural bend or break in the scree plots.[Cattell, 1966]
- Whether there are any significant negative residual variances indicative of model misspecification. [Dillon et al., 1987]

- The model fit (RMSR, Chi-Square, and RMSEA (values below 0.08 taken as suggesting acceptable fit). As the factors are allowed to correlate we are not able to use the percentage of the variance that is accounted for by the factors.

- We also considered the ratio of the first eigenvalue to the second eigenvalue as indicative of the presence of a single general factor with additional methods factors. The larger the ratio the stronger the assumption of unidimensionality, although recommended ratios vary from 3:1 to 30:1. [van der Linden and Hambleton, 2013]

- We also looked at the quality of the variables measuring the factors and considered the size of the loadings (> 0.32 considered to be loading to a factor) and any cross-loadings (0.32 on more than one item) with a view to understanding the conceptual meaning of the factors and considering whether the variables that loaded on a factor share some conceptual meaning?.[Tabachnick and Fidell, 2001] Correlation between factors is reported; strong correlations (rho≥0.5) may indicate that factors are tapping into the same latent construct.

2.4.3.2. *Confirmatory Factor Analysis (CFA)*

CFA was then run to test a number of alternative models. These included 1) the most promising model(s) from the EFA; 2) A purely theoretically driven model based on intended theoretical constructs to be identified by the instruments/items, including appropriate correlations between items. Theoretical constructs were based on a conceptual assessment of the measures [Peasgood et al., 2014] as well as in the literature; 3) A bifactor model [Gibbons & Hedeker, 1992] in which a general factor(s) is modelled plus additional nuisance or methods/instrument factors. This allows a separation of the instrument/methods effects to help reveal the underlying structure of latent factors.

Bifactor models had a general subjective well-being factor and instrument/measurement factors, with the general and instrument/methods factors assumed to be uncorrelated. Within this bifactor approach each item response is considered to be a function of the general factor (or factors) and one specific secondary factor, which is orthogonal to the general factor. The secondary factors account for the residual variance (after that accounted for the general factor) shared by a set of items. The secondary factors could be a common construct which is only partly picked up in the general factor (such as a sub-theme), or it could be a nuisance factor such as shared measurement response (such as the use of negative wording or shared response options).

Within the CFA cross-loaders were included only in their main factor (except for the bifactor models), and separate item error correlations were only included where it would not be sensible to include another factor within the model and it has a theoretical justification (e.g. shared measurement error arising due to similar phrasing or response options). Main factors were allowed to correlate.

The MPlus software produces 'Modification Indices' (MI) that make suggestions about changing the model in order to improve overall model fit. Modification Indices show how the chi squared would reduce if a factor/item was to be freed (the constraint removed)[6]. However, MI are sensitive to sample size. We considered the MI cautiously, focusing on those of 10 or above, and potential factor loadings above 0.3, but only where there was a theoretical justification to support a change to the model. [Schreiber et al, 2006]

The best CFA model was identified judged partly on model performance criteria including:

- **Chi Squared**: Chi-squared test indicates the difference between the observed and expected covariance matrices. Ideally this should be insignificant, however, due to the large sample sizes this is unlikely to arise. Within nested models an improved model should see a significant reduction in the Chi-squared.
- **RMSEA** (Root Mean Square Error of Approximation): This measures whether an *a priori* model reproduces observed data patterns. The RMSEA ranges from 0 to 1 with values under 0.06 considered to be good model fit.[Hu and Bentler, 1999] A good fitting model should have a 90% confidence interval for the RMSEA below 0.08 on the upper bound. The RMSEA penalises overly complex models.
- **CFI** (Comparative Fit Index): This analyses whether the hypothesised model fits the data better than a more restricted baseline model, while adjusting for the issues of sample size. It ranges from 0 to 1 with values above 0.95 considered good fit, and values above 0.90 adequate. [Hu and Bentler, 1999]
- **TLI** (Tucker-Lewis index): As with CFI the TLI compares the hypothesised model to a restricted baseline model. The TLI ranges from 0 to 1 with values above 0.95 considered good fit, and above 0.90 adequate. [Hu and Bentler, 1999] The TLI also penalises overly complex models.

In addition to exploring model fit, the models were also critically judged on "substantive and conceptual grounds" [Morgan et al, 2015: p15].

[6] MI are equivalent to chi squared difference with single degree of freedom of a nested model. A MI of 3.84 indicate a critical value of chi squared at $p<0.05$ with 1 degree of freedom

2.5. Methods summary

HIPO was a patient dataset whereas SYC65, MIC, USoc and HSE were general population datasets with self-reported conditions. HIPO did not have a healthy group so the MIC healthy group (n=1,472) was added to this data to allow comparisons with a 'no condition' group. Table 1 summarises the measures in each dataset and the direct comparisons that can be performed. HIPO, SYC65 and MIC allowed comparisons of ONS-4[7] and the health measures; USoc and HSE allowed comparison of GHQ-12 and S-WEMWBS and one of the 2 health measures. ICECAP-A was available in MIC while ICECAP-O was available in SYC65. None of the datasets covered all the measures and ONS-4 and ICECAP-A could not be compared to either WEMWBS or GHQ-12 but there were a number of other single item SWB questions included in the analysis in all the datasets.

Table 1: Direct comparisons across datasets

	ONS-4	S-WEMWBS	GHQ-12	ICECAP-A	ICECAP-O	ASCOT	EQ-5D
S-WEMWBS	x						
GHQ-12	x	USoc, HSE					
ICECAP-A	MIC		x				
ICECAP-O	SYC65	SYC65					
ASCOT	SYC65	SYC65	x	SYC65			
EQ-5D	HIPO,MIC	HSE	HSE	MIC	SYC65	SYC65	
SF-6D	HIPO,MIC	USoc	USoc	MIC		x	HIPO, MIC

Table 2 summarises the analysis that was undertaken to address the three questions of interest. Summary scores and distributions provide some indication that measures may be different, which informs the first two questions. Correlation analysis, effect sizes and factor analysis were used to provide information for all 3 questions. Strong correlations indicate similarities in measures. Factor analysis provided information on whether or not items (positive and negative as well as health) come from a single underlying trait or domain. Relative effect sizes of 1 provide evidence that SWB measures assessed differences in the same way as health measures. Relative effect sizes greater than 1 indicated that SWB were more sensitive, while less than 1 indicates that they were less sensitive than the health measures.

[7] Note that there was a difference in two of the ONS-4 questions for the three datasets – happy and anxious, which refer to 'today' in HIPO and SYC65 refer to 'yesterday' in MIC.

Table 2: analysis undertaken to address each question

Question	Compare scores	Examine distributions	Correlation	Factor analysis	Effect sizes
Are the SWB instruments measuring the same thing?	√	√	√	√	√
Are both positive and negative items required?	√	√	√	√	√
How do the SWB and health instruments compare?			√	√	√

3. RESULTS

3.1. Summary statistics and distributions

Mean and median SWB scores were generally high. Mean ONS-4 scores were 6 and above on a 0 to 10 point scale, as was the happiness item in the HSE, while the life satisfaction item in USoc had a mean of 5 (1 to 7 scale) (Table 3). The levels of missing data were all less than 5% for the ONS-4 (NB: there is no missing data in MIC).

Table 3: Well-being and health measures summary statistics (HIPO, MIC and SYC65)

	mean	SD	min	max	median	25 per.	75 per.	% at floor	% at ceiling	% missing
HIPO n=5,344										n=6,351
Life satisfaction	6.67	2.51	0	10	7	5	9	2.84	10.01	1.67
Worthwhile	7.18	2.50	0	10	8	6	9	1.96	17.63	2.06
Happy	7.06	2.51	0	10	8	5	9	2.06	15.18	1.59
Anxious (recoded)	7.51	2.83	0	10	9	5	10	1.83	39.93	1.62
ONS-4 total	28.42	9.19	0	40	31	23	36	0.49	0.60	2.96
ONS-4 positive	20.92	7.09	0	30	23	17	26	0.86	7.26	2.55
Positive HIPO SWB total	48.79	15.37	0	70	52	38	61	0.06	4.12	5.95
Negative HIPO SWB total	28.78	8.70	0	40	31	23	36	0.22	6.29	2.47
EQ-5D-5L	0.69	0.28	-0.594	1	0.74	0.57	0.85	0.04	19.65	3.28
SF-6D (SF-12)	0.70	0.16	0.345	1	0.67	0.57	0.86	0.88	3.31	7.84
EQ-VAS	69.72	22.50	0	100	75	55	90	0.24	3.82	1.35
SWB-VAS	69.94	23.93	0	100	75	50	90	0.58	4.90	2.65
MIC n=6808										
Life satisfaction	6.03	2.69	0	10	7	4	8	4.29	4.51	NA
Worthwhile	6.55	2.46	0	10	7	5	8	2.22	8.15	NA
Happy	6.40	2.72	0	10	7	5	9	4.32	9.62	NA
Anxious (recoded)	6.35	2.91	0	10	6	4	9	2.44	20.21	NA
ONS-4 total	25.33	8.63	0	40	26	20	32	0.44	1.70	NA
ONS-4 positive	18.98	7.17	0	30	20	15	25	1.01	2.78	NA
Pos. SWB total	22.04	4.54	6	30	23	19	25	0.00	2.39	NA
Neg. SWB total	22.95	4.73	8	31	24	20	27	0.00	1.72	NA
ICECAP-A	0.81	0.18	-0.001[ϕ]	1	0.88	0.71	0.95	0.13	11.99	NA
EQ-5D-5L	0.73	0.23	-0.51	1	0.77	0.65	0.85	0.00	18.11	NA
SF-6D (SF-12)	0.73	0.15	0.345	1	0.72	0.62	0.86	0.00	2.79	NA
SF-6D (SF-36)	0.70	0.14	0.301	1	0.70	0.61	0.81	0.00	2.78	NA

SYC65 n=1593										n=1749
Life satisfaction	7.67	1.81	0	10	8	7	9	0.44	13.12	0.29
Worthwhile	7.90	1.82	0	10	8	7	9	0.38	18.96	0.29
Happy	7.76	2.02	0	10	8	7	9	0.88	18.71	0.23
Anxious (recoded)	7.30	2.61	0	10	8	6	10	1.00	27.18	0.40
ONS-4 total	30.63	6.87	0	40	32	27	36	0.19	7.72	0.46
ONS-4 positive total	23.33	5.24	0	30	24	21	27			
Positive SYC65	46.70	10.81	0	60	49	42	55	0.06	7.22	1.32
S-WEMWBS	26.69	4.66	7	35	27	24	30	0.06	3.95	2.17
WEMWBS	52.70	9.39	16	70	53	47	59	0.00	1.19	4.06
ICECAP-O	0.83	0.13	0	1	0.87	0.79	0.91	0.13	3.83	1.43
ASCOT	0.89	0.12	-0.03	1	0.94	0.86	0.96	0.00	9.35	2.06
EQ-5D-5L	0.73	0.21	-0.33	1	0.76	0.66	0.84	0.00	14.50	1.14
EQ-VAS	76.51	19.34	0	100	80	70	90	0.19	4.39	1.77
Well-being VAS	76.46	18.16	0	100	80	70	90	0.25	3.89	0.74

per. Percentile Φ The lowest ICECAP-A value should be 0 – this negative value is a rounding error as a result of the algorithm
Floor: Minimum possible score for each measure; Ceiling: Maximum possible score for each measure
Positive score: Sum of all positive SWB items in HIPO data; sum of all positive AQoL item in MIC data; sum of all positive items in SYC65 data
Negative score: Sum of all negative SWB items in HIPO data; sum of all negative AQoL item in MIC data
NB: MIC data did not have missing data as those who did not complete the survey were removed from the data

There was evidence of ceiling effects (i.e. a large proportion at the best score indicating high levels of SWB) in the single life satisfaction, happiness, worthwhileness and anxiety questions in HIPO and SYC65. Fewer subjects are at the ceiling in the MIC dataset (see Appendix Figures 1 to 3). The distribution of the ONS-4 within MIC shows a peak at the response of 5, which is likely to have arisen due to the addition of the word 'neutral' in the middle of the response scale for these four items (Appendix Figure 2). This suggests caution should be taken when making direct comparisons to the ONS-4 responses in HIPO, particularly in relation to the anxiety question as 'neutral' anxiety has an ambiguous meaning.

Mean GHQ positive scores in USoc and HSE were around 6 and mean GHQ negative scores were around 4.5 (Table 4). GHQ-12 scores reflect a large number of 'same as usual' responses for positive items (USoc: 47%; HSE: 58%) and 'not at all' or 'no more than usual' responses for negative items (USoc: 61%; HSE: 69%). S-WEMWBS and WEMWBS scores were 25-26 and 51-52 respectively in the USoc, HSE and SYC65 (Tables 3 and 4). ICECAP and ASCOT scores were also high >0.8 There was a high level of missing data (≈16-20%) in USoc for

the SWB measures which was much higher than the level of missing data for health (0.52%). However, the health questions were asked in the interview whereas the SWB questions were self-complete which makes it difficult to compare them.

GHQ-12 overall scores and WEMWBS had lower proportions at the top end of the scale but GHQ negative had relatively high proportions at the ceiling (see Appendix Figures 3 to 5). A lower ceiling effect would be expected where a larger number of questions are asked within the measure compared with the single item questions; although in the GHQ-12, the response options of 'same as usual' and 'no more than usual' skew the distributions in a different way.

Table 4: Well-being and health measures summary statistics (USoc-wave 1 and HSE)

	mean	SD	min	max	median	25th per.	75th per.	% at floor	% at ceiling	% missing
USoc n= 37,602										n= 47,732
GHQ score ϕ	11.02	5.32	0	36	10	7	13	0.13ǂ	0.28 ǂ	16.83
GHQ positive ϕ	6.36	2.18	0	18	6	6	7	0.19 ǂ	0.51 ǂ	16.33
GHQ negative ϕ	4.66	3.64	0	18	4	2	6	0.44 ǂ	11.04 ǂ	16.38
S-WEMWBS	25.18	4.53	7	35	26	22	28	0.30	2.31	19.56
Life satisfaction	5.26	1.45	1	7	6	5	6	2.49	13.8	17.12
SF-6D (SF-12)	0.80	0.14	0.345	1	0.859	0.681	0.922	0.13	6.02	0.52
HSE n=5,709										n=7,671
GHQ score ϕ	10.76	4.63	0	36	10	7	12	0.04 ǂ	0.19 ǂ	2.59
GHQ positive ϕ	6.31	1.81	0	18	6	6	6	0.09 ǂ	0.25 ǂ	2.19
GHQ negative ϕ	4.44	3.31	0	18	4	2	6	0.18 ǂ	11.26 ǂ	1.92
SWEMWBS	25.83	4.50	7	35	26	23	29	0.19	2.42	6.62
WEMWBS	51.19	9.01	14	70	52	46	57	0.09	1.12	9.23
Happy	7.96	1.65	1	10	8	7	9	0.26	19.15	15.64
EQ-5D-3L	0.86	0.22	-0.54	1	1	0.80	1	0	55.47	4.42
EQ-VAS	79.19	16.07	0	100	80	70	90	0.02	5.38	6.99

Per. Percentile Floor: Minimum possible score for each measure i.e. low well-being Ceiling: Maximum possible score for each measure i.e. high well-being

ϕ GHQ scores – high scores indicate poor well-being

ǂ Ceiling and floor effects in GHQ-score are reversed as high scores represent low well-being

There were differences in the way respondents completed positive and negative SWB questions, with respondents less likely to report having negative SWB. There was a negative skew for all the ONS-4 items and the aggregate HIPO positive and negative SWB scores but less for the aggregate AQoL scores (MIC). There was a slight positive skew in the GHQ-12 scores which was largely driven by the large positive skew[8] in the GHQ-negative score, while the majority of the respondents reported being 'same as usual' in the GHQ-positive score which explains the differences in the means for the two scores. This may indicate that the positive and negative items are tapping into the same construct where most people are doing OK in the positive items and also do not have problems in the negative items. WEMWBS scores were generally normally distributed.

3.2. Correlations

In each dataset, the ONS-4 positive questions (life satisfaction, worthwhile and happy) were strongly correlated with each other (ρ≥0. 7) as well as with the positive aggregate scores from other items (ρ≥0.65) (Table 5). The positive ONS-4 items had strong correlations with both ICECAP-O/A (ρ≥0.58) and WEMWBS (ρ≥0.66) (Table 5). These results may reflect potential overlap between feeling satisfied, happy, having a worthwhile life and capabilities, with the strongest overlap between life satisfaction and happiness and slightly less overlap with capabilities. Correlations ranged from moderate to strong between the positive ONS-4 items and the anxious question (Table 5), which may reflect either a difference in the distribution of the anxious question where a large proportion had no anxiety and/or the possibility that this is tapping into a separate dimension. Correlations between the positive ONS-4 items were larger with the positive aggregate SWB items (ρ=0.65 to 0.85) compared to the negative aggregate SWB items (ρ=0.54 to 0.70).

The GHQ-12 score had strong correlations with S-WEMWBS and WEMWBS (ρ≥-0.61) but slightly lower correlations with life satisfaction (USoc: ρ=-0.49) or happiness (HSE: ρ=-0.59) (Table 6). Surprisingly, correlations were lower between the S-WEMWBS and the GHQ-positive sub-score compared to S-WEMWBS and the GHQ-negative sub-score (USoc: ρ -0.50 vs. -0.59; HSE: ρ -0.51 vs. -0.63). This was also the case with life satisfaction (USoc: ρ= -0.40 vs. -0.48) and happiness (HSE: ρ= -0.46 vs. -0.58). The lack of variation in the GHQ positive sub-score, discussed in the previous section, from the response options of the GHQ-12 (most respondents report 'same as usual') may explain this result. Life satisfaction and happiness had borderline moderate correlations with WEMWBS (Table 6). Overall, correlations between GHQ-12 and S-WEMWBS were strong but smaller in absolute size compared to the ONS-4 correlations, and not of a magnitude which would imply they are tapping into the same construct. The questions covered by both measures cover qualitatively similar concepts (see Appendix 2 for the measures) and are both asked in terms of frequency, but the response options of the GHQ-12 means that the relationship between the two measures is weaker than expected.

[8] As the GHQ is coded such that a high score represents poor SWB, this is equivalent to a negative skew in the other SWB scales.

Table 5: Spearman correlations between well-being, health and social care measures (HIPO, MIC and SYC65)

HIPO n =5,344	satisfaction	worthwhile	happy	recoded anxious	ONS-4 total	ONS-4 positive total	Positive HIPO SWB total	Negative HIPO SWB total
Life satisfaction	1.00							
Worthwhile	0.80	1.00						
Happy	0.84	0.80	1.00					
Anxious (recoded)	0.60	0.56	0.67	1.00				
ONS-4 total	0.91	0.88	0.93	0.80	1.00			
ONS-4 positive total	0.94	0.92	0.94	0.65	0.97	1.00		
Positive total	0.84	0.80	0.85	0.64	0.88	0.89	1.00	
Negative total	0.67	0.64	0.70	0.70	0.77	0.72	0.72	1.00
EQ-5D-5L	0.63	0.52	0.56	0.47	0.62	0.61	0.63	0.55
SF-6D (SF-12)	0.72	0.62	0.68	0.58	0.74	0.72	0.73	0.67
EQ-VAS	0.70	0.58	0.63	0.48	0.67	0.68	0.67	0.57
SWB-VAS	0.82	0.73	0.79	0.61	0.83	0.83	0.84	0.68

MIC n =6,808	satisfaction	worthwhile	happy	recoded anxious	ONS-4 total	ONS-4 positive total	Positive AQoL SWB total	Negative AQoL SWB total	ICECAP-A
Life satisfaction	1.00								
Worthwhile	0.77	1.00							
Happy	0.76	0.72	1.00						
Anxious (recoded)	0.32	0.28	0.40	1.00					
ONS-4 total	0.87	0.82	0.88	0.64	1.00				
ONS-4 positive total	0.93	0.89	0.91	0.37	0.94	1.00			
Positive total	0.69	0.64	0.66	0.35	0.72	0.73	1.00		
Negative total	0.62	0.55	0.62	0.43	0.69	0.66	0.82	1.00	
ICECAP-A	0.65	0.58	0.60	0.31	0.65	0.67	0.81	0.74	1.00
EQ-5D-5L	0.39	0.31	0.35	0.24	0.40	0.39	0.53	0.55	0.56
SF-6D (SF-12)	0.50	0.43	0.48	0.36	0.56	0.52	0.65	0.69	0.65
SF-6D (SF-36)	0.48	0.41	0.46	0.36	0.54	0.50	0.63	0.67	0.64

SYC65 n=1593	satisfaction	worthwhile	happy	recoded anxious	ONS-4 total	ONS-4 positive total	Positive SWB total	S-WEMWBS	WEMWBS	ICECAP-O
Life satisfaction	1.00									
Worthwhile	0.75	1.00								
Happy	0.80	0.71	1.00							
Anxious (recoded)	0.50	0.43	0.57	1.00						
ONS-4 total	0.85	0.80	0.88	0.81	1.00					
ONS-4 positive total	0.92	0.89	0.92	0.55	0.92	1.00				
Positive total	0.79	0.76	0.77	0.50	0.79	0.84	1.00			
SWEMWBs	0.66	0.68	0.69	0.52	0.73	0.73	0.78	1.00		
WEMWBS	0.68	0.69	0.71	0.51	0.74	0.75	0.80	0.96	1.00	
ICECAP-O	0.63	0.59	0.60	0.47	0.65	0.65	0.71	0.68	0.70	1.00
EQ-5D-5L	0.44	0.36	0.40	0.37	0.47	0.43	0.49	0.43	0.46	0.47
ASCOT	0.48	0.44	0.45	0.34	0.49	0.50	0.54	0.49	0.51	0.56
EQ-VAS	0.59	0.50	0.54	0.41	0.59	0.59	0.64	0.55	0.59	0.57
WB-VAS	0.74	0.65	0.70	0.50	0.74	0.76	0.76	0.68	0.71	0.69

strong: ≥|0.5|, moderate: <|0.5 to ≥0.3| and weak: <|0.3 to ≥0.1| [Cohen 1992]

Table 6: Spearman correlations between well-being scores and health (USoc wave 1 and HSE)

USoc n =37,602	GHQ score	GHQ positive	GHQ negative	S-WEMWBS score	Life satisfaction
GHQ score ϕ	1				
GHQ positive ϕ	0.86	1			
GHQ negative ϕ	0.95	0.65	1		
S-WEMWBS score	-0.61	-0.50	-0.59	1	
Life satisfaction	-0.49	-0.40	-0.48	0.50	1
SF-6D (SF-12)	-0.56	-0.47	-0.54	0.42	0.36

HSE n = 5,709	GHQ score	GHQ positive	GHQ negative	S-WEMWBS score	WEMWBS score	Happy
GHQ score ϕ	1					
GHQ positive ϕ	0.82	1				
GHQ negative ϕ	0.95	0.60	1			
SWEMWBS score	-0.64	-0.50	-0.62	1		
WEMWBS score	-0.66	-0.51	-0.63	0.96	1	
Happy	-0.59	-0.46	-0.58	0.56	0.56	1
EQ-5D-3L	-0.46	-0.45	-0.40	0.36	0.39	0.34

strong: ≥|0.5|, moderate: <|0.5 to ≥0.3| and weak: <|0.3 to ≥0.1| [Cohen 1992]
ϕ GHQ scores – high scores indicate poor well-being

The correlations between SWB and health measures were generally lower for EQ-5D (ρ= 0.24 to 0.63, small to large) than between SWB measures themselves (Tables 5 and 6). Correlations were slightly higher between SF-6D and the SWB measures (ρ= 0.36 to 0.74, Tables 5 and 6). In HIPO and SYC65, the correlations were moderate to large between the SWB measures and EQ-VAS (ρ=0.41 to 0.70) compared to large between SWB measures and the WB-VAS (ρ= 0.50 to 0.84). The weakest correlations across all the measures were with the ONS-4 anxious question in HIPO, MIC and SYC65, the WEMWBS and life satisfaction/happiness in USoc and HSE. SWB measures had moderate to large correlations with ASCOT (SYC65: ρ= 0.34 to 0.56) indicating that it is measuring something different from SWB. ASCOT also had moderate correlations with the EQ-5D (ρ= 0.47). The moderate to large correlations between SWB measures and health and social-care measures indicates that although there is overlap, these measures are not capturing the same thing.

3.3. Factor analysis

A number of models were tested for each dataset; the models and model performance are reported separately for each dataset in this section. Models where Chi-square is not statistically significant, RMSEA has a 90% confidence interval for the RMSEA below 0.08 on the upper bound, CFI and TLI above 0.95 were considered good (CFI and TLI above 0.90 were adequate).

HIPO data

Factor analysis based on HIPO data used items from the SF-12, EQ-5D-5L, ONS-4 ONS-mood questions (angry, bored, content, lonely) various SWB evaluative items (WB-VAS, life going well, looking forward to tomorrow) and flourishing items (supportive social relationships, think a lot of happiness of others, capable (can do things I want to do), spend most of time doing things I enjoy).

Initially physical health items were included into the EFA to identify wellbeing items which might map closely to physical health. The question on feeling 'capable' loaded to physical health, however, this was kept in initially for the CFA since it has a conceptual link to wellbeing being linked to concepts of autonomy. The two SF-12 items on emotional problems impacting on role functioning and work or other activities, also formed a separate factor. As these two items would not be able to form a robust factor in CFA they were dropped from the analysis.

The EFA of mental health and well-being items identified a possible two factor solution with items separating based on positive or negative wording, and some items (happy, content and bored) as cross-loaders. The ratio of first to second Eigenvalues was 10.390:1.170 (8.88) and this was indicative of a single main factor.

A number of CFA models were tested (see Table 7). All models excluded the main items that identified physical health. CFA models included a model with all the items (Model 1), then models split based on wording (Models 2 and 3), a theoretical construct model (Model 4) and bifactor models (Models 2 and 6). Model fit statistics indicated that a model with two negatively worded factors (Model 3), one for anxiety/depression items and the other for negative mood showed some improvement compared to a single negative factor. The purely theoretical separation (Model 4) did not fit the data well. The bifactor model gave the best fit, particularly when capable (which in the EFA loaded into a physical health factor) was excluded (Model 6).

Table 7: Confirmatory factor analysis of mental health and well-being items (HIPO)

Model	Items	Chi Squared (df)	RMSEA (95% CI)	CFI	TLI	Corr.
1 [all]	All well-being questions (excluding physical health and role functioning)	13793.620 (135) p=0.000	0.126 (0.125-0.128) 0.000	0.529	0.466	
2 [positive; negative]	Factor A: Positively worded well-being questions Factor B: Negatively worded well-being questions	6906.882 (df 134) p=0.000	0.089 (0.087-0.091) 0.000	0.766	0.733	
3 [positive; negative; negative affect]	Factor A: Positively worded well-being questions Factor B: Negatively worded well-being questions depression and anxiety (EQ-5D anxiety/depression, SF-12 calm and peaceful and SF-12 downhearted and blue. Factor C: Negative affect or feelings (lonely, bored, angry)	5670.083 (df 132) p=0.000	0.081 (0.080-0.083) 0.000	0.809	0.778	
4 [life evaluation; mood; social activities]	Factor A: Questions about life evaluation (life satisfaction, WB-VAS, life going well, look forward to tomorrow) Factor B: Questions about mood (happy, content, bored, angry, anxious, anxiety/depression, downhearted) [additional correlations between bored, anxious and angry] Factor C: Questions about activities and social relations (worthwhile activities, lonely, care about happiness of others, supportive social relationships, spend most of time doing things I enjoy, capable)	9787.960 (df 129) p=0.000	0.109 (0.107-0.110)	0.667	0.605	
5 [bifactor model: general factor; negative affect; negative mental health; positive]	Factor A: General SWB Factor B: Negative mood questions Factor C: Negative mental health questions Factor D: Positive questions	2919.393 (df 116) p=0.000	0.062 (0.060-0.064) 0.000	0.903	0.872	B-C -0.452
6 [bifactor model: general factor without capable; negative affect; negative mental health; positive]	Factor A: General SWB, excluding capable Factor B: Negative mood questions Factor C: Negative mental health questions Factor D: Positive questions, excluding capable	2269.320 (df 101) p = 0.000	0.058 (0.056-0.060)	0.923	0.896	B-C -0.446

MIC data

MIC data contained items from ICECAP-A, AQoL, ONS-4 and SF-36. The ONS variable 'anxious' was not included in the analysis because the inclusion of the term 'neutral' to the response option results in response distributions being very differently to other variables. The items within the SF-36 that ask about both physical and/or emotional health impacting on social activities (SF-36 question 20, and question 32) were also excluded since this could be referring to either physical or emotional issues. It also worth noting that some of the AQoL questions style and response spans the positive-negative spectrum meaning they cannot be easily classed as a positively or negatively worded item. These include:

- Calm-Agitated (When you think about whether you are calm and tranquil or agitated, are you…: Always calm and tranquil to always agitated)
- Close relationships (Your close relationships (family and friends) are…: Very satisfying to very unpleasant)
- Enjoy close relationships (How much do you enjoy your close relationships (family and friends)?..: immensely to I hate them)
- How much do you feel you can cope with life's problems?...: completely to not at all

The initial EFA identified a clear physical health factors (all pain items, all mobility and physical functioning items, hearing, vision, self-care, independence, being a burden, and sleep) which were excluded from the CFA. A factor was identified for emotional health impacting upon work activities which was also excluded from the CFA since it was likely to pick up whether an individual was currently employed in addition to functioning.

A three factor model was identified by EFA for remaining well-being items (excluding physical health and work-related items): Social relationships and enjoyment; depression and anxiety; and energy. However, many items cross loaded. The ratio of first to second Eigen values (22.228:1.902) (11.69) was suggestive of a strong single well-being factor.

As with HIPO, several models were tested (see Table 8). All items combined (Model 1) did not fit the data well. An improvement was found when positively and negatively worded items were allowed to separate (Model 2). Relying on theoretical differences to drive the factors fitted the data better (Model 3), but a very high correlation was found between negative items (low mood versus agitated mood (0.959). These were combined in Model 4 with marginal improvements in model fit. An additional factor was included in Model 5 to separate out flourishing items (control, coping, settled, achievement, worthwhile). However, this did not improve model fit and the factor was highly correlated with the first, positively worded factor.

The best fitting model was a bifactor model which had three main factors (all mood, evaluation and flourishing items; social relationships; energy) and three methods factors (positively worded items; negatively worded items; ICECAP items). The model fit was reasonable.

Table 8: Confirmatory factor analysis of mental health and well-being items (MIC)

Model	Items	Chi squared (df) P	RMSEA	CFI	TLI	Correlations
1 (1 factor)	All items	65951.27 (629) P= 0.000	0.124 (0.124-0.125)	0.890	0.884	na
2 (2 factors)	Factor A: All positively worded items Factor B: All negatively worded items	52927.23 (628) p= 0.000	0.111 (0.110-0.111)	0.912	0.907	A-B 0.905
3 (5 factors)	Factor A: positively worded items (evaluation, mood and flourishing) Factor B: social relationships (friendship (ICE), enjoy relationships (AQoL), Intimate relationship (AQoL), satisfaction relationships (AQoL) and isolation (AQoL), excluded (AQoL)) Factor C: negatively worded items on mood (depression and sad) Factor D: agitated/calm (calm, worry, anxiety) Factor E: energy (worn out, energy (SF36 questions 23 and 27), tired, enthusiastic)	45724.86 (656) p= 0.000	0.100 (0.100-0.101)	0.926	0.921	A-B 0.869; A-C 0.913 A-D 0.894; A-E 0.850 B-C 0.815; B-D 0.743 B-E 0.703; C-D 0.959 C-E 0.801; D-E 0.826
4 (4 factors)	Factor A: positively worded items (evaluation, mood and flourishing) Factor B: social relationships Factor C: negatively worded items on mood (depression and sad) and agitated/calm (calm, worry, anxiety) Factor D: energy	43410.60 (623) p= 0.000	0.100 (0.100-0.101)	0.929	0.924	A-B 0.869; A-C 0.915 A-D 0.850; B-C 0.805 B-D 0.703; C-D 0.814
5 (5 factors – separate flourishing items factor)	Factor A: positively worded items (evaluation and mood) Factor B: social relationships Factor C: negatively worded items on mood (depression and sad). agitated/calm (calm, worry, anxiety) Factor D: energy Factor E: positively worded flourishing items (control, coping, worthwhile)	49841.819 (df 620) p= 0.000	0.108 (0.107-0.109)	0.918	0.912	A-B 0.879; A-C 0.904 A-D 0.804; A-E 0.949 B-C 0.805; B-D 0.659 B-E 0.822; E-C 0.904

6 (Bifactor: 3 factors and 3 methods /nuisance factors)	General factors: A: positively and negatively worded items (evaluation, affect, flourishing) B: energy (worn out, energy (SF-36 questions 23 and 27), tired, enthusiastic) C: social relationships (friends and isolation) Methods/instrument factors: D: positively worded item on feelings, evaluation and energy E: negatively wording items on feelings F: ICECAP questions	28232.047 (df 593) p=0.000	0.083 (0.082-0.084)	0.954	0.948	A-B 0.817 A-C 0.857 B-C 0.682 D-E 0.192 D- F 0.244 E-F -0.035

SYC65 data

The SYC65 analysis was based on items from WEMWBS, ONS-4, ASCOT, ICECAP-O, and additional wellbeing items (capable, looking forward, support relations, contributes to happiness of others, enjoy activities, life going well). The items in the EQ-5D on physical health (EQ1-4) and the general health question and physical health question clearly loaded into a physical health factor. The item in WEMWBS which asks about having 'energy to spare' loaded with physical health items as did ASCOT4 (safe and secure), and did not cross-load with other factors. However, these were kept within the CFA since they are conceptually close to well-being. Although energy could be taken as a physical symptom the item here is taken from a well-being instrument (the WEMWBS). The questions in ASCOT (8 and 9) that relate to how being helped makes people feel in relation to dignity loaded clearly onto a separate factor and is conceptually different to overall subjective well-being. These items were excluded from the CFA as there were only 2 items which would not make it possible to identify a separate factor.

The EFA on the well-being items for the SYC65 identified a five factor solution (after excluding those items that only load to physical health):

- General WB (including ICECAP (enjoy) EQ-VAS and WB-VAS)
- WEMWBS items (except loved, and interested in new things) and ICECAP (future), EQ-D anxious/depressed
- ICECAP items (except loved and future) and ASCOT items
- Relationships/loved ICECAP (Question on love), WEMWBS questions 9 and 12 (close to others, feeling loved)
- Interested new things (WEMWBS), anxiety/depression (EQ-5D)

As expected the factors split predominantly via instrument rather than theoretical construct. However, the items on relationships and love appear to form a distinct factor to the overall well-being items. A possible bifactor model was implied with first factor Eigen value of 20.405 and second 2.524 (ratio 8.12). All the items from this EFA with the exception of EQ-VAS were included in the CFA.

The first CFA combined model (Model 1) did not fit the data well (see Table 9). A very small improvement was found when separating out the 3 items from the ASCOT on food and cleanliness (Model 2). This factor only correlated at 0.684 with the remaining factor suggesting it was picking up something different to the main well-being items. Adding three other factors for relationships, clarity, and control further improved the model fit, although the latter two remain highly correlated with the main well-being factor (Model 3).

A bifactor model provided the best fit for the data giving an adequate model fit (Model 4). This model had separate instrument/nuisance factors for each instrument (0-10 scale items, ASCOT, WEMWBS, and ICECAP). The ASCOT and ICECAP instrument factors (i.e. the residual part that is not picked up by the main factor) were allowed to correlate since they both ask questions framed as capability.

Table 9: Confirmatory factor analysis of mental health and well-being items (SYC65)

Model	Items	Chi Squared of Model fit (df)	RMSEA (90% CI)	CFI	TLI	Correlations
1	All items (excluding EQ-VAS)	12309.74 (df 665) p =0.000	0.100 (0.099-0.102)	0.798	0.786	
2	Factor A: all items except food and cleanliness Factor B: ASCOT food and cleanliness	11987.39 (df 664) p =0.000	0.099 (0.097-0.100)	0.802	0.792	A-B 0.684
3	Based on theoretical distinctions and the results of the EFA Factor A: main SWB Factor B: ASCOT food and cleanliness Factor C: relationships/loved Factor D: clarity Factor E: control The highest residual variance is in anxious (ONS)	9155.16 (df 655) P =0.000	0.086 (0.085-0.088)	0.852	0.841	A-B 0.652 A-C 0.842 A-D 0.895 A-E 0.916 B-C 0.558 B-D 0.580 B-E 0.847 C-D 0.747 C-E 0.662 D-E 0.784
4	Bifactor model : General factors Factor A: main SWB Factor B: ASCOT food and cleanliness Factor C: relationships/loved Methods factors Factor D: ASCOT items Factor E: WEMWBS items Factor F: 0-10 scale items and ONS items Factor G: ICECAP items	5740.86 (df 628) p=0.000	0.068(0.067-0.070)	0.911	0.901	A-B 0.785 A-C 0.815 B-C 0.626 G(ICECAP)- D(ASCOT) 0.715

HSE data

The analysis for HSE contained items from the GHQ-12, EQ-5D, WEMWBS and a single happy item. Physical health items loaded onto a single physical health factor and these were excluded. EFA of the remaining items favoured a 3 factor solution: GHQ negative items (and anxiety from the EQ-5D); GHQ positive items; and the WEMWBS items along with the single 'happy' item. The EFA separates items such as "been losing confidence in yourself?" (GHQ) from "I've been feeling confident" and "I've been feeling useful" from "felt that you are playing a useful part in things", which lacks credibility and suggests the EFA is driven by instrument effects. A bifactor structure was indicated by the ratio of first to second eigenvalue value 13.145:2.60 (5.06).

In the CFA five models were tested (see Table 10). The single factor model had poor fit (Model 1). The model suggested by the EFA (Model 2), which had positive and negative factors by instrument, had better fit than simply splitting the items by positive/negative wording (Model 3). A theoretical split model (confidence, strain, concentration, positive emotions, and relationships with others) did not fit the data well with evidence of high correlation between factors (Model 4). The best fitting model was a bifactor model in which all items were included within a main factor with 3 methods factors for the positive/negative GHQ and the WEMWBS.

Table 10: Confirmatory factor analysis of mental health and well-being items (HSE)

Model	Items	Chi Squared of Model fit (df)	RMSEA (90% CI)	CFI	TLI	Correlations
1	All items	51200.54 (df 351) p=0.000	0.137 (0.136-0.138)	0.828	0.815	
2	3 factors as identified in the EFA ▪ Factor A: GHQ positive items ▪ Factor B: GHQ negative items and anxiety (EQ-5D) ▪ Factor C: WE and the happy item	17914.88 (df 347) p=0.000	0.081 (0.080 0.082)	0.941	0.935	A-B 0.759 A-C 0.627 B-C 0.737
3	2 factors: all positive and all negative ▪ Factor A: GHQ positive and WEMWBS and the happy items ▪ Factor B: GHQ negative and anxiety (EQ-5D)	31471.78 (df 349) p=0.000	0.108 (0.107 0.109)	0.895	0.886	A-B 0.798
4	4 factors based on theoretical differences in the items (drawing on Graetz, 1991): ▪ Factor A: confidence: (feeling loved, feeling good about self, worthless, confidence (GHQ), confident (WE)) ▪ Factor B: strain: (dealing with problems well (WE), lost sleep over worry (GHQ), able to face problems (GHQ), constantly under strain (GHQ), couldn't overcome difficulties (GHQ), relaxed (WE), energy to spare (WE) interested in new things (WE) ▪ Factor C: concentrate: felt capable of making decisions (GHQ), able to concentrate (GHQ), been able to make up my own mind about things (WE), been thinking clearly (WE) ▪ Factor D: happy: been feeling cheerful (WE), optimistic about future (WE), able to enjoy day-to-day activities (GHQ), reasonably happy (GHQ), unhappy (GHQ) ▪ Factor E: other people: interested in others (WE), playing useful part in things (WE), feeling close to others (WE), feeling useful (WE)	37971.69 (df 340) p=0.000	0.120 (0.119 0.121)	0.873	0.858	A-B 0.923 A-C 0.895 A-D 0.962 A-E 0.908

5	Bifactor model: General factor ▪ A: All items Methods Factors ▪ B: GHQ positive items ▪ C: GHQ negative items ▪ D: WE items	15136.26 (df 321) p =0.000	0.078 (0.077-0.079	0.950	0.941	B-D 0.074 C-D -0.081 B-C - 0.179

WE = WEMWBS

USoc data

Analysis in Understanding Society used items from the GHQ-12, S-WEMWBS, life satisfaction, and SF-12. This was similar to HSE in terms of the mental health/well-being measures apart from that the shorter (7 item) version of the WEMWBS was included.

As with the other datasets, initial EFA with physical health items indicated a clear physical health factor and a factor which linked to role functioning that may be picking up employment (items SF-12 on emotional problems impacting on activities, questions 6 and 7). As with the other datasets these were excluded. The EFA without health items indicated a three factor solution: GHQ positive; negative items; and WEMWBS items with life satisfaction which mirrored the HSE results. A possible bifactor structure was indicated by the high ratio of first to second eigenvalues (10.067:1.946) (5.17).

Four models were tested in the CFA (see Table 11); the model with positive and negative wording was not tested. As with HSE results, the single factor model (Model 1) and the theoretical factor model (Model 3) had poor fit and there was evidence of high correlation in the theoretical model. The factor drawing on the EFA results (Model 2) had adequate fit and as with the other datasets, the bifactor model (Model 4) with methods/instrument factors provided the best fit to the data.

Table 11: Confirmatory factor analysis of mental health and well-being items (USoc)

Model	Items	Chi Squared of Model fit (df)	RMSEA (90% CI)	CFI	TLI	Corr.
1	All items	141201.13 (df 209 p = 0.000	0.131 (0.130-0.132)	0.842	0.826	
2	Drawing on EFA – based on instruments Factor A: WEMWBS items and life satisfaction Factor B: positive GHQ and calm/peaceful (SF-12 question 11) Factor C: negative GHQ and downhearted (SF-12 question 9)	67901.16 (df 206) p = 0.000	0.091 (0.091-0.092)	0.924	0.915	A-C 0.725 B-C 0.839 A-B 0.660
3	Theoretical model as for HSE Factor A: concentration Factor B: strain and low mood Factor C: happy and confident Factor D: relationships to others	129019.64 (df 203 p=0.000	0.127 (0.127 0.128)	0.856	0.836	A-B 0.945 A-C 0.796 A-D 0.914 B-C 0.947 B-D 0.821 C-D 0.828
4	Bifactor model General factor A: All items (WEMWBS, life satisfaction, calm and peaceful (SF-12 question 11), downward and blue (SF-question 9), GHQ-12) Methods factors B: WEMWBS items C: GHQ negative items D: GHQ positive items Correlations of main factor to instrument factors held at 0.	43790.32 (df 187 p=0.000	0.077 (0.076-0.078)	0.951	0.940	B-C 0.231 B-D 0.259 C-D 0.511

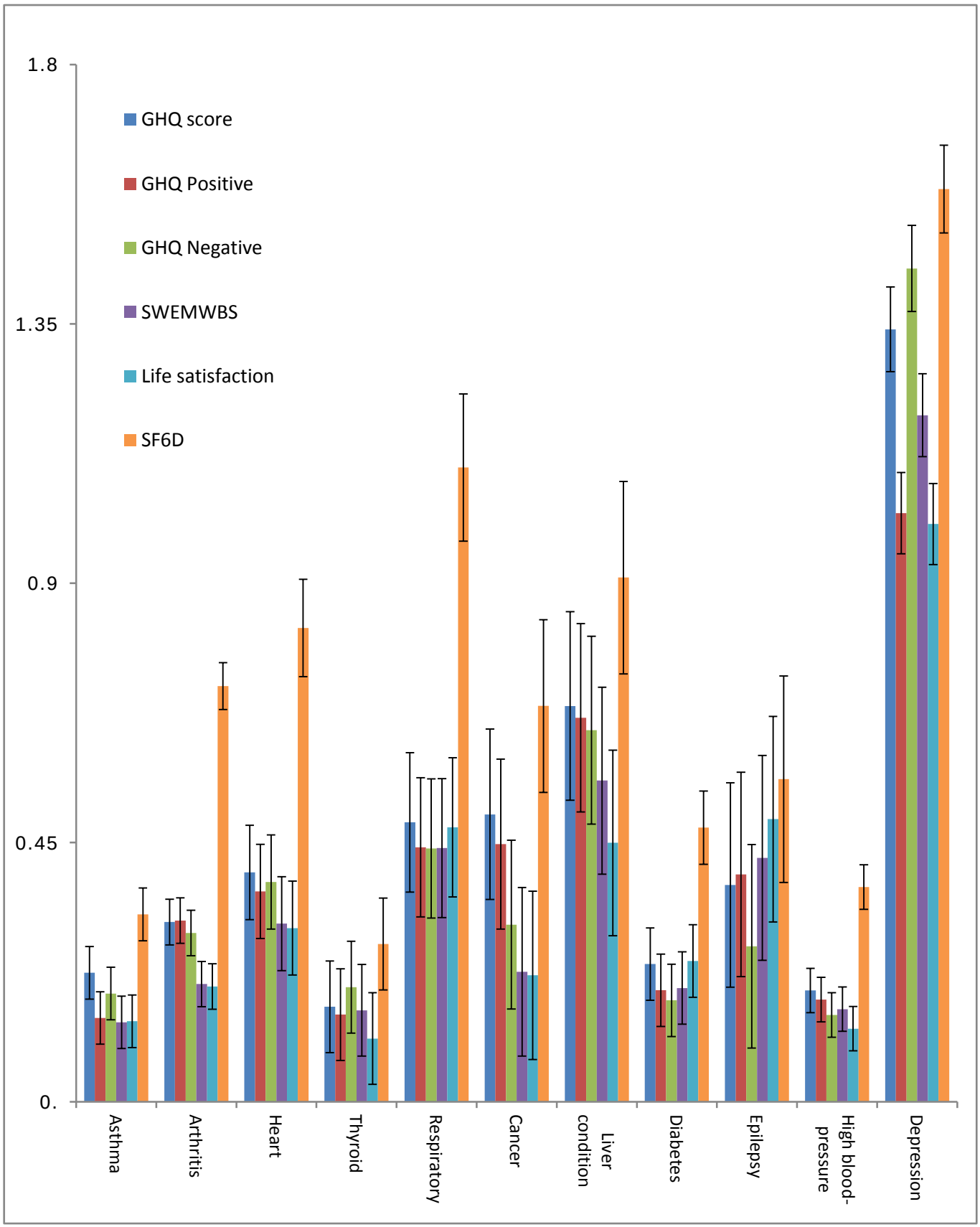
An initial exploratory factor analysis conducted on the wellbeing and health items in each dataset found that factors tended to reflect different instruments rather than underlying theoretical constructs. This suggested that the response style, order effects (which may be particularly problematic in large questionnaires and where items are placed in a grid structure) and response options (or instrument/method effects) may have been dominating the factor structure.

The more sophisticated factor analysis that could separate the instrument/method effects from the underlying latent factors.

3.4. Effect sizes

Across the five datasets, the Eta^2 for the all the SWB measures and health measures were small (see Appendix Tables 3 to 7 which show the full regressions). However, this does not mean that there was no effect, using the mean based Cohen's d effect size Graph 1 shows SWB and SF-6D effect sizes which range from 0.1 to over 1. We focus on the Eta^2 effect sizes taken from the regression analysis, however, as this allows a more accurate control for other important covariates.

Figure 1: Cohen's d effect size for health conditions with 95% CI (USoc wave 1)



In the HIPO data the SWB Eta^2 effect sizes ranged from 1 to 31% of those for the EQ-5D-5L for different physical health conditions (Table 12). There were no mental health conditions in HIPO. There were some differences across the three positive ONS-4 items but not in a systematic way. The anxious ONS-4 question was not included in the analysis due to lack of a suitable comparator. SWB measures had a larger effect sizes for unemployment relative to EQ-5D-5L (Relative effect: 1.39 to 2.47), see Appendix Table 3.

Table 12: Relative effect sizes – SWB Eta^2/ EQ-5D Eta^2 (HIPO)

	satisfaction	worthwhile	happy	ONS-4 total	EQ-5D-5L
Relative to EQ-5D-5L					
Certain Infections	0.07	0.03	0.02	0.00	1.00
Neoplasms	0.07	0.26	0.09	0.26	1.00
Blood Disorders	0.24	0.30	0.21	0.00	1.00
Endocrine And Metabolic	0.03	0.03	0.05	0.22	1.00
Nervous System	0.10	0.05	0.08	0.01	1.00
Circulatory	0.03	0.00	0.03	0.02	1.00
Respiratory	0.06	0.02	0.05	0.01	1.00
Digestive	0.00	0.05	0.01	0.07	1.00
Skin	0.06	0.02	0.01	0.31	1.00
Musculoskeletal	0.04	0.01	0.03	0.00	1.00
Genitourinary	0.09	0.02	0.06	0.04	1.00
Unclassified signs and symptoms	0.13	0.04	0.14	0.01	1.00
External Causes	0.07	0.01	0.06	0.00	1.00
Other Factors	0.03	0.21	0.04	0.31	1.00
Miscellaneous	0.14	0.05	0.04	0.01	1.00

Φ Healthy group from MIC data; Relative effect sizes shown where EQ-5D is statistically significant at the 1% level.
NB: Known group analysis was not undertaken for the ONS-4 anxious question and the aggregate HIPO positive and negative scores due to the lack of appropriate comparators in the MIC data. The ONS-4 anxious question in MIC had a 'neutral' label and this may have affected responses (see Appendix Figure 2).

In MIC, the effect sizes for the positive ONS-4 items ranged from 6 to 38% of the EQ-5D-5L effect sizes for the physical health conditions and from 49 to 61% for depression, with life satisfaction having the largest effect sizes (Table 13). ICECAP-A and the positive AQoL aggregate SWB had larger effect sizes than the ONS-4 positive items with relative effect sizes ranging from 20 to 100% of the EQ-5D-5L in physical health and around 133% to 134% for

depression indicating that depression explained more of the variation in these measures than in EQ-5D-5L. The anxious question from ONS-4 did not outperform the positive items or EQ-5D-5L in either the physical or mental health conditions (1% to 34%) while the negative AQoL SWB total did better than the positive ONS-4 items for physical health and better than the EQ-5D-5L for mental health (Table 13).

Table 13: Relative effect sizes – SWB Eta^2/ EQ-5D Eta^2 (MIC)

	satisfaction	worthwhile	happy	anxious (recoded)	ONS-4 total	ICECAP-A	positive AQoL SWB total	negative AQoL SWB total	EQ-5D-5L
Relative to EQ-5D-5L									
Asthma	0.21	0.14	0.26	0.06	0.23	0.52	0.58	0.44	1.00
Cancer	0.38	0.23	0.36	0.14	0.40	0.59	0.53	0.46	1.00
COPD	0.25	0.13	0.21	0.15	0.28	0.33	0.38	0.40	1.00
Depression	0.55	0.49	0.61	0.31	0.73	1.33	1.34	1.09	1.00
Diabetes	0.35	0.29	0.28	0.10	0.36	0.58	0.56	0.42	1.00
Hearing	0.34	0.23	0.37	0.34	0.48	0.85	1.00	0.51	1.00
Arthritis	0.12	0.06	0.06	0.01	0.08	0.23	0.20	0.19	1.00
Heart	0.32	0.21	0.25	0.11	0.32	0.52	0.53	0.39	1.00

Relative effect sizes calculated where EQ-5D is statistically significant.

In SYC65, the effect sizes for the positive ONS-4 items ranged from 2 to 108% in the physical health conditions compared to EQ-5D-5L effect sizes (Table 14a). The presence of cancer was able to explain more of the variation in happiness compared to EQ-5D-5L (108%). WEMWBS, S-WEMWBS and ICECAP-O had effect sizes that were 4 to 68% of EQ-5D-5L effect sizes for the physical health conditions. All the positive SWB measures effect sizes were more than 1.5 times the effect sizes of EQ-5D-5L for depression but only ICECAP-O performed slightly better than EQ-5D-5L for other mental health conditions (Table 14a). The anxious ONS-4 question did not outperform any of the positive measures for all the physical health conditions and depression but did better for other mental health conditions as the effect size was 1.35 times that of EQ-5D-5L. ASCOT effect sizes ranged from 6 to 88% of EQ-5D-5L in the physical and other mental health conditions but were 2.78 times larger for depression. Results based on change indicated none of the measures out-performed EQ-5D-5L for new physical conditions but the S-WEMBWS and ASCOT outperformed EQ-5D-5L for new mental health conditions (Table 14b) but these results should be interpreted with caution due to small sample size.

Table 14a: Relative effect sizes – SWB Eta^2/ EQ-5D Eta^2 (SYC65)

	satisfaction	worthwhile	happy	anxious (recoded)	ONS-4 total	ONS-4 positive total	S-WEMWBS	WEMWBS	ICECAP-O	ASCOT	EQ-5D-5L
Relative to EQ-5D-5L											
Arthritis	0.07	0.04	0.02	0.07	0.07	0.05	0.04	0.04	0.04	0.06	1.00
Cancer	0.88	0.67	1.08	0.08	0.77	1.33	0.27	0.59	0.33	0.72	1.00
COPD	0.35	0.36	0.19	0.02	0.24	0.33	0.27	0.24	0.28	0.50	1.00
Depression	1.83	1.70	2.52	1.35	2.59	1.73	1.76	2.03	2.57	2.78	1.00
Diabetes	0.72	0.34	0.58	0.03	0.45	0.49	0.43	0.52	0.68	0.88	1.00
Heart disease	0.29	0.26	0.32	0.01	0.24	0.26	0.39	0.54	0.32	0.44	1.00
Stroke	0.30	0.20	0.00	0.10	0.16	0.15	0.26	0.10	0.31	0.60	1.00
Other mental health	0.16	0.21	0.10	1.26	0.55	0.31	0.88	0.55	1.04	0.20	1.00
Other physical health	0.31	0.06	0.18	0.00	0.13	0.21	0.13	0.17	0.24	0.28	1.00

Relative effect sizes calculated where EQ-5D is statistically significant.

Table 14b: Relative effect sizes – SWB Eta^2/ EQ-5D Eta^2 (SYC65 – Change follow-up - baseline)

	satisfaction	worthwhile	happy	anxious (recoded)	ONS-4 total	ONS-4 positive total	S-WEMWBS	WEMWBS	ICECAP-O	ASCOT	EQ-5D-5L
Relative to EQ-5D-5L											
New Mental health	0.04	0.05	0.16	0.00	0.02	0.03	1.64	0.62	0.06	1.52	1.00
New Physical health	0.20	0.61	0.47	0.12	0.15	0.58	0.16	0.00	0.10	0.02	1.00

In USoc, a similar pattern was observed for GHQ, S-WEMWBS and life satisfaction as with the other SWB measures with effect sizes that were 1 to 92% of the SF-6D effect sizes for physical health conditions but larger effect sizes for clinical depression for the GHQ (Table 15a). The negative GHQ performed better than the positive GHQ but this was due to the lack of variation in the positive GHQ items already noted in section 3.1. There was follow-up data in USoc which allowed assessment of change. Only three conditions had statistically significant change in the SF-6D at follow-up: arthritis, cancer and clinical depression. For arthritis and cancer the effect sizes were 7 to 22% and 3 to 37% of the SF-6D, respectively (Table 15b). For clinical depression, effect sizes were 2.92 to 4.57 times that of the SF-6D for GHQ and S-WEMWBS but 0.92 times for life satisfaction. SWB measures all had relatively larger effect sizes for unemployment compared to SF-6D (Relative effect: 1.2 to 2).

Table 15a: Relative effect sizes – SWB Eta^2/ SF-6D Eta^2 (USoc wave 1)

	GHQ	GHQ positive	GHQ negative	S-WEMWBS	Life satisfaction	SF-6D (SF-12)
Relative to SF-6D						
Asthma	0.19	0.12	0.19	0.10	0.14	1.00
Arthritis	0.16	0.12	0.14	0.05	0.06	1.00
Heart problems	0.20	0.24	0.13	0.13	0.07	1.00
Hyperthyroidism (over-active thyroid)	0.09	0.01	0.15	0.71	0.24	1.00
Hypothyroidism (under-active thyroid)	0.29	0.11	0.35	0.02	0.15	1.00
Respiratory problems	0.24	0.19	0.20	0.16	0.40	1.00
Any kind of liver condition	0.78	0.92	0.52	0.41	0.13	1.00
Cancer or malignancy	0.62	0.78	0.39	0.16	0.22	1.00
Diabetes	0.10	0.10	0.07	0.15	0.41	1.00
Epilepsy	0.21	0.20	0.16	0.44	0.39	1.00
High blood pressure	0.21	0.16	0.19	0.27	0.12	1.00
Clinical depression	1.36	1.03	1.21	0.85	0.57	1.00

Relative effect sizes calculated where SF-6D is statistically significant.

Table 15b: Effect sizes - Eta^2 (USoc waves 1 and 4: score in wave 4 minus score in wave 1)

	GHQ change	GHQ positive change	GHQ negative change	S-WEMWBS change	Life satisfaction change	SF-6D (SF-12) change
Relative to SF-6D						
New arthritis	0.10	0.11	0.07	0.08	0.22	1.00
New cancer or malignancy	0.30	0.37	0.17	0.03	0.17	1.00
New clinical depression	4.37	4.52	2.92	3.23	0.91	1.00

Relative effect sizes calculated where SF-6D is statistically significant.

The USoc data also contains information on the presence or absence of a list of disabilities. Cross section analysis of these for wave 1 shows a similar picture with most disabilities showing a lower effect size relative to the that for the SF-6D (Table 15c). Memory and concentration is an exception here, showing a higher effect sizes in the GHQ measures and S-WEMWBS relative to the SF-6D, although not for life satisfaction. Disabilities of speech and being able to recognise danger are not shown as they have an insignificant SF-6D effect size (the denominator). These results are in line with the fixed effects results where the GHQ measures and S-WEMWBS had relatively higher effect sizes for memory and concentration than SF-6D (Table 15d).

Table 15c: Relative effect sizes – SWB Eta^2/ SF-6D Eta^2 (USoc wave 1)

	GHQ	GHQ positive	GHQ negative	WEMWBS	Life satisfaction	SF-6D (SF-12)
Relative to SF-6D						
Mobility	0.07	0.07	0.05	0.06	0.10	1.00
Carrying	0.12	0.08	0.12	0.06	0.06	1.00
Dexterity	0.24	0.15	0.23	0.02	0.25	1.00
Continence	0.39	0.22	0.40	0.37	0.20	1.00
Hearing	0.00	0.10	0.08	0.13	0.02	1.00
Sight	0.32	0.13	0.36	0.12	0.10	1.00
Memory or concentration	1.99	1.86	1.51	1.29	0.71	1.00
Balance	0.36	0.57	0.18	0.15	0.05	1.00
Personal care	0.53	0.78	0.28	0.19	0.30	1.00
Other disability	0.46	0.33	0.40	0.25	0.25	1.00

Relative effect sizes calculated where SF-6D is statistically significant.

Table 15d: Effect sizes – Cohen's f^2 (USoc waves 1 and 4: fixed effects -deviation from individual level mean)

	GHQ	GHQ positive	GHQ negative	S-WEMWBS	Life satisfaction	SF-6D (SF-12)
Relative to SF-6D						
Mobility	0.04	0.06	0.02	0.05	0.08	1
Carrying	0.06	0.04	0.05	0.02	0.07	1
Dexterity	0.66	0.53	0.53	0.26	0.42	1
Continence	0.07	0.36	0.00	0.43	0.00	1
Memory or concentration	2.30	2.60	1.41	1.20	0.47	1
Balance	1.14	1.23	0.72	0.12	0.03	1
Personal care	0.81	1.06	0.43	0.15	0.28	1
Other disability	0.83	0.69	0.66	0.24	0.28	1

Relative effect sizes calculated where SF-6D is statistically significant.

In HSE, the SWB effect sizes were also relatively smaller than the EQ-5D-3L ones for the physical health conditions with the exception of conditions of the genitourinary system and other complaints for GHQ and GHQ positive (Table 16). The SWB effect sizes were relatively larger than EQ-5D-3L effect sizes including for the single happiness question (Table 16). Although the effect sizes for unemployment were relatively larger for most of the SWB measures than EQ-5D-3L, they were only statistically significant for the GHQ and GHQ negative.

Table 16: Relative effect sizes – SWB Eta^2/ EQ-5D Eta^2 (HSE)

	GHQ	GHQ positive	GHQ negative	S-WEMWBS	WEMWBS	Happy	EQ-5D-3L
Relative to EQ-5D-3L							
Neoplasms and benign growths	0.68	0.72	0.48	0.23	0.29	0.15	1.00
Endocrine and metabolic	0.04	0.01	0.05	0.23	0.31	0.15	1.00
Mental health disorders	2.05	1.54	1.75	1.52	1.56	1.23	1.00
Nervous system	0.21	0.26	0.13	0.24	0.26	0.18	1.00
Heart and circulatory system	0.28	0.20	0.24	0.40	0.51	0.16	1.00
Digestive system	0.32	0.17	0.32	0.09	0.13	0.21	1.00
Genitourinary system	1.02	1.42	0.58	0.82	0.95	0.66	1.00
Musculoskeletal condition	0.11	0.10	0.08	0.08	0.09	0.06	1.00
Other complaints and infectious disease	1.41	2.23	0.71	0.75	0.72	0.39	1.00

Relative effect sizes calculated where EQ-5D is statistically significant.

4. DISCUSSION AND RECOMMENDATIONS

This report summarises psychometric and factor analysis which sought to compare well-being measures that are used or recommended for use in the UK, including the ONS-4 (life satisfaction, worthwhileness, happiness and anxious), WEMWBS and S-WEMWBS, GHQ-12 and ICECAP-A/O. Analysis assessed the relationship between these measures as well as in relation to health measures (EQ-5D and SF-6D) and social care measures (ASCOT). Factor analysis was used to assess whether or not the measures covered more than one dimension and if they were separate from health. Furthermore, the relative ability of the SWB measures to discriminate between groups with known differences compared to health measures was examined. Five large datasets covering a number of self-reported or hospital diagnoses condition groups were used to inform the analysis. The aim of the analysis was to address three related questions. Are SWB measures identifying different constructs? Is it necessary to include positive and negative items in a measure of SWB? What would be the potential impact of using SWB measures rather than health measures to evaluate health care?

4.1. Summary and discussion of findings

4.1.1. Are SWB measures identifying the same constructs?

Overall, the results suggest that the SWB measures were closely related, particularly when comparing positive SWB items/measures. Assessment of the performance of the positive SWB items/scores/sub-scores in known group analysis indicated that generally they performed in a similar way. The 'happy' and 'worthwhile' questions in the ONS-4 are highly correlated with 'life satisfaction' in HIPO, MIC and SYC65. There were also strong correlations between these items and other measures such as ICECAP and WEMWBS. These latter correlations were not as strong as the correlations between the ONS-4 positive items but there was still considerable overlap. This is quite surprising because a greater distinction between measures of positive affect and measures of evaluation was anticipated. Conceptually they differ and considerable empirical evidence has often found them to both form separate factors and correlate differently to other variables.[e.g. Arthaud-Day et al, 2005; Kahneman and Deaton, 2010] It is possible that differences between the three ONS positive questions are being disguised by shared measurement error because they are completed in the same way immediately after each other.

Factor analysis, particularly the bifactor CFA models, indicated that evaluative questions and positive affect questions were more likely to be in a single factor. Separate constructs only emerged for items related to social relationships, energy and other objective wellbeing items.

4.1.2. Do we need negative items in addition to positive items?

There was less overlap between the positive and negative items, in line with linguistic and conceptual differences between the absence of happiness and the presence of sadness. The distribution of scores for negative items indicated that there were fewer respondents overall reporting problems in the negative SWB items. This would imply that:

a. either positive and negative affect (and/or experiences or psychological capacities) are different dimensions of a life. If so, data about one need not necessarily provide information about the other.
b. or they share the same underlying construct but have different measurement error, for example, people may be more willing to share information about experiencing low levels of positive concepts than they are about admitting negative experiences.

Although EFA identified separate positive and negative factors, the more sophisticated CFA suggested that these were largely instrument effects rather than differences in positive and negative SWB. In all of the six datasets the best performing model is a bifactor model with methods/instrument factors modelled in addition to either a single well-being factor, or additional factors relating to social relationships, energy and food/cleanliness. This suggests that positive and negative affect (or experiences or psychological capacities) fall into the same underlying well-being dimension. The improved fit of the bifactor model suggests the different factors in the EFA, which clearly identifies separate GHQ-12 and the S-WEMWBS factors for example, could be driven by differences in instrument layout and response options rather than underlying differences in the concepts. It is worth noting that this approach is looking for models which give the best fit to the data overall, and whilst typically these positive and negative items/constructs may track each other very closely they may come apart for small but policy relevant sub-groups (such as the very elderly or those with mental health problems).

The effect size analysis found that the GHQ-12 negative score showed larger effect sizes than the S-WEMWBS which may indicate that it is capturing something different. The effect sizes for age and gender show some differences in all datasets between positive and negative questions. Analysis on the Gallup data from the US has found that the age profile of SWB measures differs according to affect with some types of negative affect (stress, anger) showing a greater improvement with age than other negative affect (sadness) and positive affect (enjoyment).[Stone et al, 2010] This points to a need for greater differentiation of negative affect in order to understand which aspects are separate from positive affect. There were also differences in the way negative and positive SWB questions were associated with unemployment, but this was not consistent between datasets with a larger positive than negative effect size for the MIC, but larger negative than positive for HSE and similar effect sizes for the USoc.

In comparing the GHQ-12 and S-WEMWBS, the analysis is limited by the response options of the GHQ-12. The inclusion of 'same as / no more than usual' response options means individuals who may have had on-going positive/negative experiences could have chosen this option. This is a problem that cannot be addressed by either caseness scoring or a corrected binary scoring as there is no way to tell which reference point respondents are using when they consider their 'usual' positive or negative states. The GHQ-12 was not available with other negative SWB items and so we cannot say definitively that the negative items are providing additional information. However, given the discussion about response options above, the GHQ-12 would never be a first choice for a measure of interpersonally comparable SWB. However, its presence allows comparability with historic data, and other surveys in which it is used, plus it can provide the number of people with possible depression or anxiety disorders. Moreover, from a policy perspective we are particularly concerned with the experiences of suffering themselves, and with individuals with low well-being.

There are a number of other factors to consider when making the overall judgement on the necessity of including both positive and negatively worded items in any measure of SWB. Where items fall into the same dimension they may still give information on different parts of the spectrum of that dimension. This could be explored further using IRT (Item response theory) to explore the contribution towards the precision in measurement at different points along the latent factor made by items. Instruments with only positive items may avoid accidental mistakes arising from people not carefully reading items and their response choices (for example, some completing the ONS-4 may accidently report high levels of anxiety after completing prior questions in which high scores equate with high well-being). It may also be the case that the inclusion of both negative and positive items could be used to identify respondents who are not engaging with the exercise (for example, respondents who always tick the response in the same position on the questionnaire). If such respondents can be identified then their inclusion in any data analysis can be explored in sensitivity analysis. Issues of face validity and acceptability within the intended users are also important. The general public may prefer to respond to positively mental health questions, yet those with moderate-severe mental health problems are likely to find negatively worded items a closer fit to their current experience.

4.1.3. What is the potential impact of using SWB to evaluate health care interventions?

As expected, health has less overlap with SWB measures than the overlap between SWB measures, with correlation analysis indicating that health and SWB measures are measuring separate constructs. Factor analysis resulted in a health factor that included physical functioning, pain and usual activities. The inclusion of some items related to capability or independence in this physical health factor may indicate the strong correlation between physical health and being able to perform everyday activities. It may also indicate that respondents think about physical health more when considering issues of being capable or independent in

the context of health surveys. However, the ICECAP-O items did not load strongly onto factors other than one related to relationships which may be due to differences in the capability concept. Those who completed the ICECAP-O were older and may have struggled to answer the questions; there is some evidence that respondents can struggle with capability questions. [Al Janabi et al, 2013] As would be expected, items from the ASCOT that were related to living standards such as food and accommodation loaded onto a separate factor. This supports the presence of a factor related more directly to physical health and aspects of need that are indirectly related to health.

Physical health contributes to SWB but only in a modest way. Effect sizes for physical health conditions were much smaller for SWB measures than for the EQ-5D and the SF-6D which was as expected. Results were mixed for depression or mental health with GHQ-12 (and GHQ negative) doing better than health measures as did some of the aggregate positive and negative SWB scores. Panel data confirmed these findings. S-WEMWBS did better than EQ-5D-3L and about the same as SF-6D, while the single item ONS-4 generally did worse than the health measures. In evaluating change, the GHQ and S-WEMWBS were better than SF-6D at discriminating where there was new depression, whereas they had lower discriminatory power for new asthma and new arthritis. Life satisfaction did not perform better than SF-6D in detecting new depression.

This data suggests that evaluative, life satisfaction type measures would still be less sensitive to changes in mental health than existing health measures; however, the relative importance of mental health compared to physical health conditions would increase substantially. Consequently, shifting to SWB measures would result in relatively lower weight for physical conditions and potentially larger weights for mental health depending on which measure was used. However, health measures may place relatively lower weight on other non-health differences of interest such as unemployment. Further panel data is required in conditions other than the three conditions with statistically significant changes in the SF-6D at follow-up in the USoc and the two conditions with statistically significant changes in EQ-5D at follow-up in the SYC65 data.

The lower sensitivity of SWB measures to physical health conditions and as such would have considerable implications for sample sizes required to detect changes in physical health.

4.2. Limitations

This study benefits from repeating the analysis on five different datasets. There were slightly different relationships emerging between the measures. This suggests the possible variability of the relationships between the measures and possible interaction with characteristics such as age which varies across the datasets. These differences between datasets offer a caution against over interpretation from a single finding.

Although the analysis benefited from the use of several large patient and general population datasets, there are a number of limitations. Complete case analysis was undertaken which excluded a large number of respondents, although the datasets were still large. Those who were excluded had lower health and SWB which may have had an impact on the results. This may be particularly problematic if the relationships between the measures are not constant across the distribution, for example two measures of SWB may look very similar for happy people, but less similar for unhappy people. Furthermore, none of the datasets had all the measures of interest which made it difficult to compare across all the measures although using several datasets allowed assessment of whether results were consistent across them which strengthens the overall findings. No data contained patients with moderate to severe mental health problems, who are an important group that may have different relationships between the measures considered here.

There were also limitations which affected the comparison of SWB and health measures. The way in which health conditions are defined varied between datasets, from self-report to self-report with restrictions to hospital allocated ICD-10 code which may have an impact on the assessment of the performance of SWB measures compared to health measures. For example, broad ICD categories may be too heterogeneous to provide information on specific conditions. In addition, only USoc and SYC65 had follow-up data but there was limited change in the latter dataset. The problem with being limited to cross-sectional data is that we were unable to assess how well measures responded to change or control for unobserved characteristics such as personality which may influence SWB. Assessing response to change is particularly important when considering the use of SWB measures to assess health conditions and interventions in the health sector.

There were differences in the mode of administration across the datasets, which may impact upon some questions (such as negative affect) more strongly than others (such as more objective health questions which may have an impact in the comparisons for the 3rd question). Research has found that on average lower scores to well-being questions are received if the interview is carried out via self-completion rather than administered by an interviewer, particularly for female respondents.[Pudney, 2010] In HSE and USoc SWB items and HRQoL items were self-compete with an interviewer present which may limit comparison with the other datasets. There were also differences in HIPO which is a patient dataset six weeks post discharge and though respondents were not as healthy as the general population, they may have had higher SWB because they had received treatment. Furthermore, the comparator group was the healthy group from MIC and these respondents answered a different questionnaire online which involved a large number of measures and there were differences in the wording of the question on happiness and anxious for ONS-4 (today vs. yesterday), as well as the inclusion of a 'neutral' label at 5 on the 0 to 10 scale.

4.3. Implications for policy

a. The results do not provide definite guidance on whether GHQ-12 and S-WEMWBS are both required. However, if the aim is to provide a measure of SWB that can be compared across individuals, then replacing the GHQ-12 should be considered due to the response options of the items. S-WEMWBS may not be sufficient due to the absence of negative items, while the ONS-4 suffer from less reliability as they are single item measures.

b. The implications of any move to using SWB to evaluate health policy needs to be carefully considered. Moving to SWB would result in a substantial increase in the weight given to mental health compared to physical health conditions.

c. SWB measures, including those focusing on psychological well-being, are far less sensitive to physical health conditions and as such would have considerable implications for sample sizes required to detect changes in health.

References

Al-Janabi H, Flynn T, Coast J. Development of a self-report measure of capability well-being for adults: the ICECAP-A. *Quality of Life Research.* 2012; 21, 167-176.

Al-Janabi H, Keeley T, Mitchell P, Coast J. (2013) Can capabilities be self-reported? A think-aloud study. *Social Science and Medicine*. 87: 116-122.

Brazier J, Usherwood T, Harper R, Thomas K. Deriving a preference-based single index from the UK SF-36 Health Survey. *Journal of clinical epidemiology*. 1998; 51(11), 1115-1128.

Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics*. 2002; 21(2), 271-292.

Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. *Medical care.* 2004; 42(9), 851-859.

Cattell RB. The scree test for the number of factors. *Multivariate Behavioral Research*. 1966; 1: 245-276.

Coast J, Flynn T, Natarajan L, Sproston K, Lewis J, Louviere J et al. Valuing the ICECAP capability index for older people. *Social Science and Medicine* 2008; 67(5):874-882.

Cohen J. *Statistical power analysis for the behavioral sciences* (2nd ed.) 1988. Hillsdale, NJ: Erlbaum.

Cohen J. Statistical power analysis. *Current directions in psychological science.* 1992; 1(3), 98-101.

Costello AB, Osborne JW. Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis. Practical Assessment, Research & Evaluation. 2005. 10 (7)

Couzner, L., Crotty, M., Norman, R. and Ratcliffe, J., 2013. A comparison of the EQ-5D-3L and ICECAP-O in an older post-acute patient population relative to the general population. Applied health economics and health policy, 11(4), pp.415-425.

Davis, J.C., Bryan, S., McLeod, R., Rogers, J., Khan, K. and Liu-Ambrose, T., 2012. Exploration of the association between quality of life, assessed by the EQ-5D and ICECAP-O, and falls risk, cognitive function and daily function, in older adults with mobility impairments. BMC geriatrics, 12(1), p.65.

Deaton A. Income, Ageing, Health and Wellbeing around the World: Evidence from the Gallup World Poll, *NBER Working Paper.* 2007; 13317.

Dillon, W. R., Kumar, A., Mulani, N. Offending estimates incovariance structure analysis: Comments on the causes of and solutions to Heywood cases. *Psychological Bulletin*. 1987. 101(1), 126-135.

Dolan P. Modelling valuation for Euroqol health states. *Medical Care.* 1997;35:351-363.

Dolan P, Metcalfe R. Measuring subjective wellbeing: recommendations on measures for use by national governments. *Journal of Social Policy*. 2012; 41 (02), 409-427.

Flynn TN, Huynh E, Peters TJ, Al-Janabi H, Moody A, Clemens S, Coast J. Scoring the ICECAP-A capability instrument. Estimation of a UK general population tariff. *Health Economics*. 2013. Early view available online.

Gibbons RD, Hedeker DR. Full-information item bi-factor analysis. Psychometrika. 1992; 57(3):423-36.

Goldberg DP, Hillier VF. A scaled version of the General Health Questionnaire. *Psychol Med*. 1979; 9:139–145.

Goldberg DP, Williams P. *A user's guide to the GHQ.* Windsor, NFER Nelson. 1988.

Goodchild ME, Duncan-Jones P. Chronicity and the General Health Questionnaire. *British Journal of Psychiatry.* 1985; 146, 55-61.

Grewal I, Lewis J, Flynn TN, Brown J, Bond J, Coast J. Developing attributes for a generic quality of life measure for older people: preferences or capabilities? *Social Science & Medicine.* 2006; 62: 1891-1901.

Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, Bonsel G and Badia X. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011; 20, 1727–36.

Holzinger KJ, Swineford F. The bi-factor method. *Psychometrika*. 1937; 2:41–54.

Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*. 1999;6(1):1-55.

Kammann R, Flett R. Affectometer 2: a scale to measure current level of general happiness. *Australian Journal of Psychology*. 1983; 82, 1007-1022.

Keeley, T., Al-Janabi, H., Nicholls, E., Foster, N.E., Jowett, S. and Coast, J., 2015. A longitudinal assessment of the responsiveness of the ICECAP-A in a randomised controlled trial of a knee pain intervention. Quality of Life Research, 24(10), pp.2319-2331.

Knabe A, Rätzel S, Schöb R, Weimann J. Dissatisfied with Life but Having a Good Day: Time-use and Well-being of the Unemployed. *Economic Journal.* 2010; 120, 867–889.

National Institute for Health and Care Excellence (NICE). Guide to the methods of technology appraisal. 2013a.

National Institute for Health and Care Excellence (NICE). The Social Care Guidance Manual. 2013b.

Netten AP, Beadle-Brown J, Caiels J, Forder JE, Malley J, Smith NJ, Windle K *et al*. ASCOT Adult Social Care Outcomes Toolkit: Main Guidance v2. 1 PSSRU Discussion Paper 2716/3. University of Kent. 2011.

Netten A, Burge P, Malley J, Potoglou D, Towers AM, Brazier J, Forder J. *et al*. Outcomes of social care for adults: developing a preference-weighted measure. *Health Technology Assessment*. 2012; 16(16), 1-166.

Morgan GB, Hodge KJ, Wells KE, Watkins MW. Are fit indices biased in favor of bi-factor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *Journal of Intelligence*. 2015;3(1):2-0.

Mitchell PM, Al-Janabi H, Richardson J, Iezzi A, Coast J. The Relative Impacts of Disease on Health Status and Capability Wellbeing: A Multi-Country Study. *PloS one*. 2015;10(12), p.e0143590.

Muthén BO. A method for studying the homogeneity of test items with respect to other relevant variables. Journal of educational statistics. 1985;10:121–132.

Peasgood T, Brazier J, Mukuria C, Rowen D. A conceptual comparison of well-being measures used in the UK. EEPRU Research Report. 2014.

Podsakoff PM, MacKenzie SB, Lee JY, Podsakoff NP. Common method biases in behavioral research: A critical review of the literature and recommended remedies. J Appl Psychol 2003, 88: 879–903.

Pudney S. *An experimental analysis of the impact of survey design on measures and models of subjective wellbeing* (No. 2010-20). ISER Working Paper Series. 2010.

Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press, 1960.

Relton C, Bissell P, Smith C, Blackburn J, Cooper CL, Nicholl J, Tod A, Copeland R, Loban A, Chater T, Thomas K, Young tm Weirt C, Anderson g, Millbourn A & Manners, R. (2011). South Yorkshire Cohort: a 'cohort trials facility' study of health and weight-Protocol for the recruitment phase. *BMC public health*, *11*(1), 640.

Richardson, J., Chen, G., Khan, M.A. and Iezzi, A., 2015. Can Multi-attribute Utility Instruments Adequately Account for Subjective Well-being?. Medical Decision Making, 35(3), pp.292-304.

Richardson J, Iezzi A, Khan MA, Maxwell A. 2013. Validity and Reliability of the Assessment of Quality of Life (AQoL)-8D Multi-Attribute Utility Instrument. *The Patient-Patient-Centered Outcomes Research.* 1-12.

Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. 2012. A practical guide to calculating Cohen's f^2, a measure of local effect size, from PROC MIXED. *Frontiers in psychology*, 3.

Schreiber JB, Nora A, Stage FK, Barlow EA, King J. Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of educational research.* 2006. 99(6): 323-338.

Stewart-Brown S, Tennant A, Tennant R, Platt S, Parkinson J, Weich S. Internal construct validity of the Warwick-Edinburgh mental well-being scale (WEMWBS): a Rasch analysis using data from the Scottish health education population survey. *Health and Quality of Life Outcomes.* 2009: 7(1), 15-22.

Stone AA, Schwartz JE, Broderick JE, Deaton A. 2010. A snapshot of the age distribution of psychological well-being in the United States. *Proceedings of the National Academy of Sciences*. 107(22), 9985-9990.

Streiner DL, Norman GR. 2008. *Health measurement scales: a practical guide to their development and use*. Oxford university press.

Tabachnick BG, Fidell LS. Using multivariate statistics.2001.

Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, Parkinson J, Secker J & Stewart-Brown S. The Warwick-Edinburgh mental well-being scale (WEMWBS): development and UK validation. *Health and Quality of life Outcomes.* 2007*;* 5(1), 63.

Teresi JA. Overview of quantitative measurement methods. Equivalence, invariance, and differential item functioning in health applications. Med Care. 2006 Nov;44(11 Suppl 3):S39-49.

Tesio L. Measuring behaviours and perceptions: Rasch analysis as a tool for rehabilitation research. Journal of Rehabilitation Medicine. 2003;35(3):105-15.

van der Linden WJ, Hambleton RK, eds. *Handbook of modern item response theory*. Springer Science & Business Media. 2013.

van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki, D, Lloyd A, Scalone L, Kind P & Pickard AS. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value in Health*. 2012; 15(5), 708-715.

van Leeuwen, K.M., Bosmans, J.E., Jansen, A.P., Hoogendijk, E.O., van Tulder, M.W., van der Horst, H.E. and Ostelo, R.W., 2015. Comparing measurement properties of the EQ-5D-3L, ICECAP-O, and ASCOT in frail older adults. Value in Health, 18(1), pp.35-43

World Health Organization. International Statistical Classification of Diseases and Related Health Problems. 10th Revision. 2010.

Young T, Yang Y, Brazier J, Tsuchiya A. The Use of Rasch Analysis in Reducing a Large Condition-Specific Instrument for Preference Valuation: The Case of Moving from AQLQ to AQL-5D. Medical Decision Making. 2011;31(1):195-210.

Youny T, Yang Y, Brazier JE, et al. The first stage of developing preference-based measures: constructing a health-state classification using Rasch analysis. Quality of Life Research. 2009 Mar;18(2):253-65.

Appendix 1

Appendix Table 1: Background characteristics (HIPO and MIC)

	HIPO n= 5,344			MIC n=6,808			SYC65 (n=1,593)	
	No.	%		No.	%		No.	%
Age (mean, s.d.)	59.2	16.39		51.2	15.11		72.64	5.77
Female	2,686	50.26		3,690	54.2		811	50.91
Employment status								
▪ Employed	2,018	37.76		3,095	45		94	5.9
▪ Unemployed	103	1.93		561	8.24		0	0
▪ Retired	2,046	38.29		1,680	24.68		1,355	85.06
▪ Long-term sick	538	10.07		509	7.48		15	0.94
▪ Other	270	5.05		738	10.84		86	5.4
▪ Missing	369	6.9		225	3.3		43	2.7
Condition								
Certain infectious and parasitic	53	0.99	None	1,472	21.62	None	424	26.8
Neoplasms	628	11.75	Asthma	726	10.66	Arthritis	593	33.9
Blood and related disorders	121	2.26	Cancer	691	10.15	Asthma	121	6.9
Endocrine, nutritional and metabolic	84	1.57	COPD	66	0.97	Cancer	81	4.6
Nervous system	153	2.86	Depression	773	11.35	Kidney disease	17	1.0
Eye and adnexa	255	4.77	Diabetes	779	11.44	COPD	72	4.1
Circulatory system	545	10.2	Hearing Problems	713	10.47	Dementia	3	0.2

Respiratory system	221	4.14	Arthritis	796	11.69	Depression	94	5.4
Digestive system	744	13.92	Heart	792	11.63	Diabetes	190	10.9
Skin and subcutaneous tissue	134	2.51				Heart disease	218	12.5
Musculoskeletal system	712	13.32				High blood pressure	504	28.8
Genitourinary system	375	7.02				Parkinson's	10	0.6
Unclassified signs and symptoms	527	9.86				Stroke	44	2.5
External causes (injury, poisoning etc.)	296	5.54				Other mental health	24	1.4
Factors influencing health status	301	5.63				Other physical health	391	22.4
Miscellaneous ICD Chapters	78	1.46				Missing	18	1.0
Missing	117	2.19						

Appendix Table 2: Background characteristics (USoc and HSE)

	USoc wave 1 n=37,602			HSE n=5,709	
	No.	%		No.	%
Age (mean, s.d.)	45.3	17.75		47.7	17.83
Female	21,047	55.97		3,175	55.61
Employment status					
Employed	20,768	55.24		3,305	57.89
Unemployed	2,373	6.31		305	5.34
Retired	7,301	19.42		1,251	21.91
Long-term sick	1,253	3.33		N/A	N/A
Other	5,904	15.70		838	14.68
Missing	3	0.01		10	0.18
Condition (Ever experienced)			**'Has long standing illness'**		
Asthma	5,109	13.59	Neoplasms and benign growths	133	2.33
Arthritis	5,111	13.59	Endocrine and metabolic	481	8.43
Congestive heart failure	186	0.49	Mental disorders	214	3.75
Coronary heart disease	636	1.69	Nervous system	215	3.77
Angina	964	2.56	Eye complaints	97	1.70
Heart attack or myocardial infarction	753	2.00	Ear complaints	120	2.10
Stroke	597	1.59	Heart and circulatory system	674	11.81
Emphysema	254	0.68	Respiratory system	499	8.74
Hyperthyroidism (over-active thyroid)	351	0.93	Digestive system	266	4.66
Hypothyroidism (under-active thyroid)	1,113	2.96	Genitourinary system	134	2.35
Chronic bronchitis	733	1.95	Skin complaints	98	1.72
Any kind of liver condition	468	1.24	Musculoskeletal condition	971	17.01
Cancer or malignancy	1,292	3.44	Infectious disease	15	0.26
Diabetes	2,051	5.45	Blood and related disorders	46	0.81
Epilepsy	393	1.05	Other complaints	17	0.30
High blood pressure	6,781	18.03	None of the above	3,286	57.56
Clinical depression	2,580	6.86			
None of the above	20,068	53.37			

Appendix Table 3: Effect sizes - eta squared % (HIPO)

	satisfaction	worthwhile	happy	ONS-4 total	EQ-5D-5L	SF-6D (SF-12)
	%	%	%	%	%	%
Healthy group$^{\phi}$						
Certain Infections	-0.02	-0.01	-0.01	-0.02	**-0.34**	**-0.36**
Neoplasms	0.04	**0.15**	0.05	**0.23**	**-0.59**	**-1.19**
Blood Disorders	**-0.19**	**-0.24**	**-0.17**	**-0.13**	**-0.80**	**-1.25**
Endocrine And Metabolic	-0.01	-0.01	-0.01	0.00	**-0.18**	**-0.39**
Nervous System	**-0.17**	-0.10	**-0.15**	-0.06	**-1.80**	**-1.84**
Eye And Adnexa	**0.16**	**0.12**	0.08	**0.26**	-0.08	**-0.28**
Circulatory	-0.05	-0.00	-0.04	-0.00	**-1.50**	**-2.94**
Respiratory	-0.08	-0.02	-0.07	-0.00	**-1.20**	**-1.39**
Digestive	0.00	0.07	0.02	**0.12**	**-1.30**	**-1.97**
Skin	0.02	0.01	0.00	0.04	**-0.27**	**-0.31**
Musculoskeletal	**-0.24**	-0.04	**-0.21**	-0.04	**-6.36**	**-6.34**
Genitourinary	**-0.11**	-0.02	-0.08	-0.01	**-1.21**	**-1.78**
Unclassified signs and symptoms	**-0.31**	-0.09	**-0.34**	**-0.16**	**-2.39**	**-2.87**
External Causes	**-0.27**	0.05	**-0.24**	**-0.13**	**-3.90**	**-3.37**
Other Factors	0.01	**0.10**	0.02	**0.15**	**-0.48**	**-0.91**
Miscellaneous	-0.07	0.03	-0.02	-0.01	**-0.50**	**-0.54**
Male	0.08	0.00	0.07	0.04	**0.17**	**0.57**
Age	**-0.16**	-0.01	**-0.21**	-0.09	-0.04	0.04
Age Squared	**0.15**	0.01	**0.30**	**0.14**	-0.01	**-0.11**
Married	**2.34**	**1.82**	**1.82**	**2.08**	0.38	**0.52**
Unemployed	**-0.57**	**-0.51**	**-0.32**	**-0.51**	-0.02	**-0.23**
Relative to EQ-5D						
Certain Infections	0.07	0.03	0.02	0.00	1.00	
Neoplasms	0.07	0.26	0.09	0.26	1.00	
Blood Disorders	0.24	0.30	0.21	0.00	1.00	
Endocrine And Metabolic	0.03	0.03	0.05	0.22	1.00	
Nervous System	0.10	0.05	0.08	0.01	1.00	
Eye And Adnexa						

Circulatory	0.03	0.00	0.03	0.02	1.00
Respiratory	0.06	0.02	0.05	0.01	1.00
Digestive	0.00	0.05	0.01	0.07	1.00
Skin	0.06	0.02	0.01	0.31	1.00
Musculoskeletal	0.04	0.01	0.03	0.00	1.00
Genitourinary	0.09	0.02	0.06	0.04	1.00
Unclassified signs and symptoms	0.13	0.04	0.14	0.01	1.00
External Causes	0.07	0.01	0.06	0.00	1.00
Other Factors	0.03	0.21	0.04	0.31	1.00
Miscellaneous	0.14	0.05	0.04	0.01	1.00

Φ Healthy group from MIC data
Figures in bold at significant at the 1% level.
Relative effect sizes calculated where EQ-5D is statistically significant.

NB: Known group analysis was not undertaken for the ONS-4 anxious question and the aggregate HIPO positive and negative scores due to the lack of appropriate comparators in the MIC data. The ONS-4 anxious question in MIC had a 'neutral' label and this may have affected responses (see Appendix Figure 2).

Appendix Table 4: Effect sizes - eta squared % (MIC)

	satisfaction	worthwhile	happy	anxious (recoded)	ONS-4 total	ICECAP-A	positive AQoL SWB total	negative AQoL SWB total	EQ-5D-5L	SF-6D (SF-12)	SF-6D (SF-36)
	%	%	%	%	%	%	%	%	%	%	%
Asthma	**-0.42**	**-0.28**	**-0.54**	**-0.12**	**-0.47**	**-0.90**	**-1.06**	**-1.17**	**-2.03**	**-1.60**	**-2.85**
Cancer	**-1.35**	**-0.80**	**-1.28**	**-0.50**	**-1.42**	**-1.63**	**-2.08**	**-1.87**	**-3.55**	**-3.77**	**-4.84**
COPD	**-0.47**	**-0.24**	**-0.40**	**-0.29**	**-0.52**	**-0.73**	**-0.61**	**-0.70**	**-1.85**	**-1.81**	**-2.01**
Depression	**-6.18**	**-5.56**	**-6.86**	**-3.49**	**-8.27**	**-12.31**	**-15.02**	**-15.08**	**-11.26**	**-13.77**	**-13.10**
Diabetes	**-1.34**	**-1.10**	**-1.07**	**-0.39**	**-1.39**	**-1.60**	**-2.21**	**-2.14**	**-3.83**	**-3.47**	**-4.58**
Hearing	**-0.32**	**-0.22**	**-0.35**	**-0.31**	**-0.45**	**-0.48**	**-0.80**	**-0.94**	**-0.93**	**-0.69**	**-1.18**
Arthritis	**-0.86**	**-0.44**	**-0.46**	-0.07	**-0.59**	**-1.39**	**-1.69**	**-1.50**	**-7.38**	**-3.64**	**-6.31**
Heart	**-1.22**	**-0.80**	**-0.93**	**-0.43**	**-1.22**	**-1.49**	**-1.97**	**-2.02**	**-3.80**	**-3.26**	**-4.69**
Male	**-0.12**	**-0.33**	**-0.08**	0.05	-0.08	0.00	-0.00	**0.54**	**0.11**	**0.33**	**0.38**
Age	**-1.09**	**-0.44**	**-0.61**	0.04	**-0.46**	**-0.95**	**-0.92**	**-0.40**	**-1.05**	**-0.09**	**-0.21**
Age squared	**1.42**	**0.71**	**0.91**	-0.00	**0.79**	**1.17**	**1.29**	**0.81**	**0.89**	**0.21**	**0.30**
Married	**3.11**	**1.78**	**2.07**	0.03	**2.01**	**2.50**	**2.00**	**0.74**	**0.54**	**0.88**	**0.59**
Unemployed	**-2.08**	**-1.44**	**-0.75**	**-0.17**	**-1.41**	**-1.04**	**-1.03**	**-0.53**	**-0.26**	**-0.46**	**-0.34**
Relative to EQ-5D-5L											
Asthma	0.21	0.14	0.26	0.06	0.23	0.52	0.58	0.44	1.00		

Cancer	0.38	0.23	0.36	0.14	0.40	0.59	0.53	0.46	1.00
COPD	0.25	0.13	0.21	0.15	0.28	0.33	0.38	0.40	1.00
Depression	0.55	0.49	0.61	0.31	0.73	1.33	1.34	1.09	1.00
Diabetes	0.35	0.29	0.28	0.10	0.36	0.58	0.56	0.42	1.00
Hearing	0.34	0.23	0.37	0.34	0.48	0.85	1.00	0.51	1.00
Arthritis	0.12	0.06	0.06	0.01	0.08	0.23	0.20	0.19	1.00
Heart	0.32	0.21	0.25	0.11	0.32	0.52	0.53	0.39	1.00

Figures in bold at significant at the 1% level.
Relative effect sizes calculated where EQ-5D is statistically significant.

Appendix Table 5: Effect sizes - eta squared % (SYC65)

	satisfaction	worthwhile	happy	anxious (recoded)	ONS-4 total	S-WEMWBS	WEMWBS	ICECAP-O	ASCOT	EQ-5D-5L
	%	%	%	%	%	%	%	%	%	%
Arthritis	**-0.76**	**-0.42**	-0.25	**-0.77**	**-0.77**	**-0.47**	**-0.49**	**-0.42**	**-0.64**	**-11.05**
Asthma	-0.04	-0.18	-0.03	-0.01	-0.07	-0.12	-0.12	-0.25	-0.06	0.05
Cancer	**-0.40**	-0.30	**-0.48**	-0.04	-0.35	-0.12	-0.26	-0.15	-0.32	**-0.45**
COPD	**-0.62**	**-0.64**	-0.34	-0.03	**-0.42**	**-0.48**	**-0.42**	**-0.50**	**-0.90**	**-1.78**
Depression	**-5.62**	**-5.23**	**-7.74**	**-4.16**	**-7.94**	**-5.39**	**-6.23**	**-7.89**	**-8.53**	**-3.07**
Diabetes	**-1.12**	**-0.53**	**-0.90**	-0.05	**-0.69**	**-0.66**	**-0.80**	**-1.04**	**-1.35**	**-1.54**
Heart disease	-0.23	-0.21	-0.25	-0.01	-0.19	-0.32	**-0.43**	**-0.26**	**-0.35**	**-0.80**
High blood pressure	-0.00	-0.01	-0.04	-0.01	-0.01	-0.00	-0.06	-0.01	-0.24	-0.16
Stroke	**-0.37**	-0.25	-0.01	-0.12	-0.20	-0.32	-0.13	**-0.39**	**-0.75**	**-1.25**
Other mental health	-0.07	-0.09	-0.04	**-0.53**	-0.23	**-0.38**	-0.23	**-0.44**	-0.09	**-0.43**
Other physical health	**-1.29**	-0.25	**-0.73**	-0.01	-0.52	**-0.53**	**-0.71**	**-1.02**	**-1.14**	**-4.15**
gender	0.06	0.01	0.08	0.02	0.03	0.02	0.00	0.03	0.10	0.02
age	**0.61**	**0.60**	**0.69**	**0.51**	**0.85**	**1.18**	**1.15**	**0.75**	**0.33**	0.16
age2	**-0.61**	**-0.60**	**-0.69**	**-0.53**	**-0.86**	**-1.22**	**-1.20**	**-0.75**	**-0.35**	-0.18
married	**0.73**	**1.22**	**0.93**	0.06	**0.80**	0.07	0.21	**1.04**	0.26	0.03
Relative to EQ-5D-5L										
Arthritis	0.07	0.04	0.02	0.07	0.07	0.04	0.04	0.04	0.06	1.00
Asthma										
Cancer	0.88	0.67	1.08	0.08	0.77	0.27	0.59	0.33	0.72	1.00
COPD	0.35	0.36	0.19	0.02	0.24	0.27	0.24	0.28	0.50	1.00

Depression	1.83	1.70	2.52	1.35	2.59	1.76	2.03	2.57	2.78	1.00
Diabetes	0.72	0.34	0.58	0.03	0.45	0.43	0.52	0.68	0.88	1.00
Heart disease	0.29	0.26	0.32	0.01	0.24	0.39	0.54	0.32	0.44	1.00
High blood pressure										
Stroke	0.30	0.20	0.00	0.10	0.16	0.26	0.10	0.31	0.60	1.00
Other mental health	0.16	0.21	0.10	1.26	0.55	0.88	0.55	1.04	0.20	1.00
Other physical health	0.31	0.06	0.18	0.00	0.13	0.13	0.17	0.24	0.28	1.00

Figures in bold at significant at the 1% level.
Relative effect sizes calculated where EQ-5D is statistically significant.

Appendix Table 6a: Effect sizes – Eta^2 % (USoc wave 1)

	GHQ $^{\phi}$	GHQ positive $^{\phi}$	GHQ negative $^{\phi}$	WEMWBS	Life satisfaction	SF-6D (SF-12)
	%	%	%	%	%	%
Asthma	**0.07**	**0.04**	**0.07**	**-0.04**	**0.05**	**-0.37**
Arthritis	**0.52**	**0.41**	**0.45**	**-0.17**	**0.19**	**-3.26**
Heart problems	**0.13**	**0.16**	**0.09**	**-0.09**	**0.05**	**-0.69**
Hyperthyroidism (over-active thyroid)	0.00	0.00	0.01	**-0.04**	0.01	**-0.05**
Hypothyroidism (under-active thyroid)	0.01	0.00	0.01	-0.00	0.00	**-0.03**
Respiratory problems	**0.11**	**0.09**	**0.09**	**-0.07**	**0.18**	**-0.45**
Any kind of liver condition	**0.10**	**0.12**	**0.07**	**-0.05**	0.02	**-0.13**
Cancer or malignancy	**0.19**	**0.24**	**0.12**	**-0.05**	**0.07**	**-0.31**
Diabetes	**0.04**	**0.04**	**0.03**	**-0.06**	**0.17**	**-0.40**
Epilepsy	**0.03**	**0.03**	**0.02**	**-0.06**	**0.05**	**-0.13**
High blood pressure	**0.09**	**0.07**	**0.08**	**-0.11**	**0.05**	**-0.40**
Clinical depression	**8.43**	**6.40**	**7.46**	**-5.25**	**3.51**	**-6.19**
Male	**-0.46**	**-0.21**	**-0.52**	0.01	**0.07**	**0.60**
Age	**0.68**	**0.42**	**0.68**	**0.09**	**-1.14**	**0.03**
Age squared	**-0.85**	**-0.30**	**-1.04**	**0.23**	**1.50**	-0.01
Married	**-0.34**	**-0.16**	**-0.37**	**0.37**	**0.89**	**0.55**
Unemployed	**0.55**	**0.41**	**0.49**	**0.50**	**-0.68**	**-0.34**
Relative to SF-6D						
Asthma	0.19	0.12	0.19	0.10	0.14	1.00
Arthritis	0.16	0.12	0.14	0.05	0.06	1.00
Heart problems	0.20	0.24	0.13	0.13	0.07	1.00
Hyperthyroidism (over-active thyroid)	0.09	0.01	0.15	0.71	0.24	1.00
Hypothyroidism (under-active thyroid)	0.29	0.11	0.35	0.02	0.15	1.00
Respiratory problems	0.24	0.19	0.20	0.16	0.40	1.00

Any kind of liver condition	0.78	0.92	0.52	0.41	0.13	1.00
Cancer or malignancy	0.62	0.78	0.39	0.16	0.22	1.00
Diabetes	0.10	0.10	0.07	0.15	0.41	1.00
Epilepsy	0.21	0.20	0.16	0.44	0.39	1.00
High blood pressure	0.21	0.16	0.19	0.27	0.12	1.00
Clinical depression	1.36	1.03	1.21	0.85	0.57	1.00

Figures in bold at significant at the 1% level.
Relative effect sizes shown only where SF-6D is statistically significant at 1% level.
ϕ GHQ scores – high scores indicate poor well-being

Appendix Table 6b: Effect sizes – Eta^2 % (USoc wave 1)

	GHQ $^{\phi}$	GHQ positive $^{\phi}$	GHQ negative $^{\phi}$	S-WEMWBS	Life satisfaction	SF-6D (SF-12)
	%	%	%	%	%	%
Mobility	**0.21**	**0.22**	**0.15**	**0.18**	**0.30**	**3.00**
Carrying	**0.37**	**0.23**	**0.35**	**0.18**	**0.17**	**2.99**
Dexterity	**0.03**	**0.02**	**0.03**	0.00	**0.03**	**0.14**
Incontinence	**0.08**	**0.05**	**0.09**	**0.08**	**0.04**	**0.21**
Hearing	0.00	0.01	0.00	0.01	0.00	**0.05**
Sight	**0.02**	0.01	**0.03**	0.01	0.01	**0.07**
Speech	**0.02**	**0.04**	0.01	0.01	**0.02**	0.00
Memory or concentration	**2.40**	**2.25**	**1.83**	**1.56**	**0.86**	**1.21**
Recognising danger	**0.02**	**0.02**	**0.02**	0.01	0.02	0.00
Balance	**0.04**	**0.07**	**0.02**	**0.02**	0.01	**0.12**
Personal Care	**0.28**	**0.42**	**0.15**	**0.10**	**0.16**	**0.53**
Other disability	**0.70**	**0.51**	**0.62**	**0.39**	**0.38**	**1.53**
Male	**0.76**	**0.34**	**0.83**	**0.03**	**0.05**	**1.15**
Age	**1.20**	**0.72**	**1.16**	**0.23**	**1.53**	**0.13**
Age squared	**1.52**	**0.60**	**1.75**	**0.49**	**2.07**	**0.11**
Married	**0.46**	**0.20**	**0.51**	**0.46**	**1.01**	**0.72**
Unemployed	**0.73**	**0.53**	**0.65**	**0.60**	**0.81**	**0.55**
Relative to SF-6D						
Mobility	0.07	0.07	0.05	0.06	0.10	1.00
Carrying	0.12	0.08	0.12	0.06	0.06	1.00
Dexterity	0.24	0.15	0.23	0.02	0.25	1.00
Incontinence	0.39	0.22	0.40	0.37	0.20	1.00

Hearing	0.00	0.10	0.08	0.13	0.02	1.00
Sight	0.32	0.13	0.36	0.12	0.10	1.00
Memory or concentration	1.99	1.86	1.51	1.29	0.71	1.00
Balance	0.36	0.57	0.18	0.15	0.05	1.00
Personal Care	0.53	0.78	0.28	0.19	0.30	1.00
Other disability	0.46	0.33	0.40	0.25	0.25	1.00

Figures in bold at significant at the 1% level.
Relative effect sizes are only shown where SF-6D is statistically significant at 5% level.
ϕ GHQ scores – high scores indicate poor well-being

Appendix Table 6c: Effect sizes – Eta^2 % (USoc waves 1 and 4: score in wave 4 minus score in wave 1)

	GHQ change ϕ	GHQ positive change ϕ	GHQ negative change ϕ	S-WEMWBS change	Life satisfaction change	SF-6D (SF-12) change
	%	%	%	%	%	%
New asthma	0.01	0.02	0.01	0.01	0.00	0.01
New arthritis	0.01	0.02	0.01	0.01	0.03	**0.14**
New stroke	0.01	0.00	0.02	**0.05**	0.00	0.01
New hyperthyroidism (over-active thyroid)	0.00	0.01	0.00	0.00	0.01	0.00
New hypothyroidism (under-active thyroid)	0.02	0.02	0.01	**0.03**	0.00	0.03
New - any kind of liver condition	**0.06**	0.05	**0.05**	0.02	0.00	0.00
New cancer or malignancy	**0.07**	**0.09**	**0.04**	0.01	**0.04**	**0.24**
New diabetes	0.00	0.00	0.01	0.00	0.00	0.00
New epilepsy	0.00	0.00	0.00	0.02	0.00	0.00
New high blood pressure	0.01	0.00	0.01	0.02	0.00	0.00
New clinical depression	**0.70**	**0.72**	**0.47**	**0.52**	**0.15**	**0.16**
New heart problems	0.02	**0.04**	0.00	0.01	0.03	0.02
New respiratory problems	0.01	0.00	0.02	0.00	0.01	0.00
Male	**0.08**	**0.06**	**0.07**	0.01	**0.04**	0.01
Age (@ wave 4)	**0.07**	0.00	**0.13**	0.01	0.01	**0.21**
Married (@ wave 4)	**0.04**	**0.04**	0.02	**0.05**	0.02	0.02
Unemployed (@wave 4)	**0.23**	**0.21**	**0.17**	**0.03**	**0.10**	0.03
Relative to SF-6D						
New arthritis	0.10	0.11	0.07	0.08	0.22	1.00
New cancer or malignancy	0.30	0.37	0.17	0.03	0.17	1.00
New clinical depression	4.37	4.52	2.92	3.23	0.91	1.00

Figures in bold at significant at the 1% level.
Relative effect sizes shown only where SF-6D is statistically significant at 1% level.
ϕ GHQ scores – high scores indicate poor well-being

Appendix Table 6d: Effect sizes – Cohen's f^2 (USoc wave 1 and wave 4 – fixed effects model)

	GHQ ϕ	GHQ positive ϕ	GHQ negative ϕ	S-WEMWBS	Life satisfaction	SF-6D (SF-12)
	%	%	%	%	%	%
mobility	0.03	**0.04**	0.02	**0.03**	**0.06**	**0.72**
carrying	0.03	0.02	0.02	0.01	**0.03**	**0.47**
dexterity	**0.05**	**0.04**	**0.04**	0.02	**0.03**	**0.08**
continence	0.00	0.01	0.00	0.01	0.00	0.02
hearing	0.02	0.01	0.01	0.00	0.01	0.00
sight	0.00	0.00	0.00	0.00	0.00	0.01
speech	**0.09**	**0.08**	**0.07**	0.02	0.01	0.02
memory	**0.49**	**0.55**	**0.30**	**0.26**	**0.10**	**0.21**
danger	**0.04**	**0.07**	0.01	0.01	0.00	0.00
balance	0.03	**0.03**	0.02	0.00	0.00	0.03
personal care	**0.17**	**0.22**	**0.09**	0.03	**0.06**	**0.20**
other	**0.18**	**0.15**	**0.14**	**0.05**	**0.06**	**0.21**
Relative to SF-6D						
mobility	0.04	0.06	0.02	0.05	0.08	1
carrying	0.06	0.04	0.05	0.02	0.07	1
dexterity	0.66	0.53	0.53	0.26	0.42	1
continence	0.07	0.36	0.00	0.43	0.00	1
memory	2.30	2.60	1.41	1.20	0.47	1
balance	1.14	1.23	0.72	0.12	0.03	1
personal care	0.81	1.06	0.43	0.15	0.28	1
other	0.83	0.69	0.66	0.24	0.28	1

Figures in bold at significant at the 1% level.
Relative effect sizes shown only where SF-6D is statistically significant at 1% level.
ϕ GHQ scores – high scores indicate poor well-being

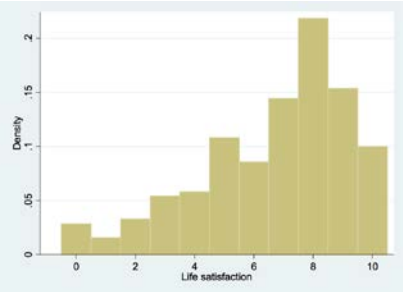
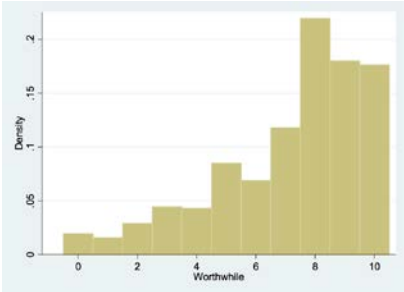
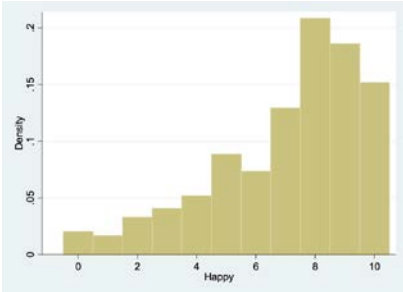
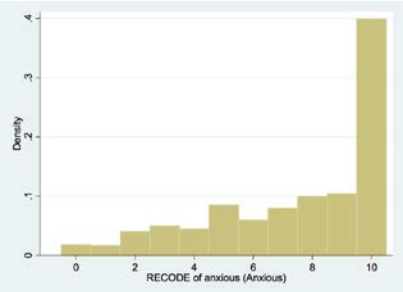
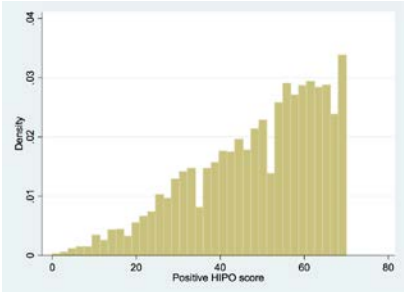
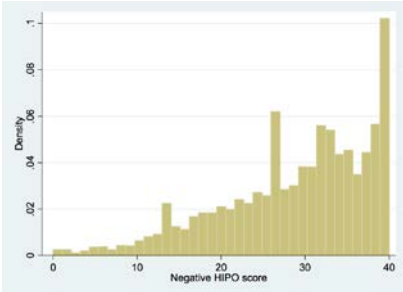
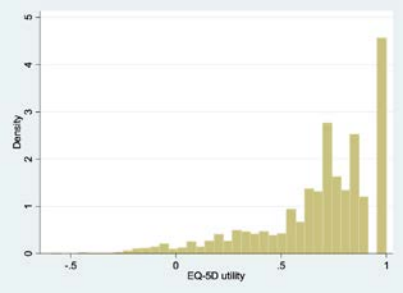
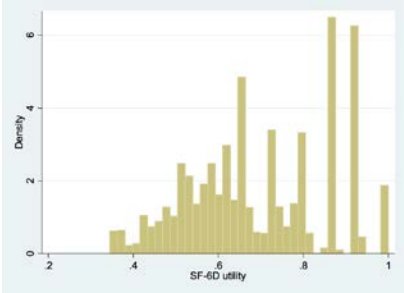
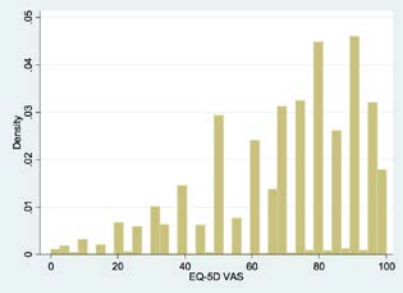
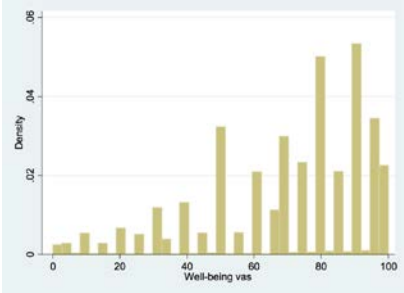
Appendix Table 7: Effect sizes – Eta^2 % (HSE)

	GHQ $^{\phi}$	GHQ positive $^{\phi}$	GHQ negative $^{\phi}$	S-WEMWBS	Full WEMWBS	Happy	EQ-5D-3L
	%	**%**	**%**	**%**	**%**	**%**	**%**
Neoplasms and benign growths	**0.44**	**0.46**	**0.31**	**-0.15**	**-0.19**	**-0.10**	**-0.64**
Endocrine and metabolic	0.01	0.00	0.01	-0.03	-0.04	-0.02	**-0.14**
Mental disorders	**5.96**	**4.49**	**5.08**	**-4.42**	**-4.53**	**-3.58**	**-2.91**
Nervous system	**0.35**	**0.45**	**0.21**	**-0.41**	**-0.44**	**-0.30**	**-1.69**
Eye complaints	0.02	-0.00	0.04	-0.04	-0.06	**-0.15**	0.00
Ear complaints	0.01	-0.00	0.02	-0.00	-0.00	0.00	-0.00
Heart and circulatory system	**0.18**	**0.13**	**0.15**	**-0.26**	**-0.33**	-0.10	**-0.64**
Respiratory system	0.08	**0.14**	0.03	-0.08	-0.08	-0.00	-0.07
Digestive system	**0.13**	0.07	**0.14**	-0.04	-0.06	-0.09	**-0.42**
Genitourinary system	**0.37**	**0.52**	**0.21**	**-0.30**	**-0.35**	**-0.24**	**-0.37**
Skin complaints	**0.12**	**0.15**	0.07	-0.03	-0.05	-0.02	-0.05
Musculoskeletal condition	**1.55**	**1.45**	**1.17**	**-1.10**	**-1.33**	**-0.89**	**-14.37**
Blood and related disorders	0.00	0.01	-0.00	0.01	0.01	-0.00	-0.06
Other complaints and infectious disease	0.11	**0.18**	**0.06**	-0.06	-0.06	-0.03	**-0.08**
Male	**-0.19**	-0.10	**-0.19**	-0.00	-0.00	**-0.10**	0.06
Age	**0.51**	**0.13**	**0.64**	0.00	-0.03	**-0.70**	-0.03
Age squared	**-0.74**	**-0.11**	**-1.05**	0.02	**0.14**	**1.02**	-0.00
Married	**-0.36**	**-0.20**	**-0.35**	**0.26**	**0.40**	**2.13**	**0.49**
Unemployed	**0.28**	0.01	**0.47**	-0.08	-0.04	-0.10	-0.03
Relative to EQ-5D-3L							
Neoplasms and benign growths	0.68	0.72	0.48	0.23	0.29	0.15	1.00
Endocrine and metabolic	0.04	0.01	0.05	0.23	0.31	0.15	1.00
Mental disorders	2.05	1.54	1.75	1.52	1.56	1.23	1.00
Nervous system	0.21	0.26	0.13	0.24	0.26	0.18	1.00
Heart and circulatory system	0.28	0.20	0.24	0.40	0.51	0.16	1.00
Digestive system	0.32	0.17	0.32	0.09	0.13	0.21	1.00

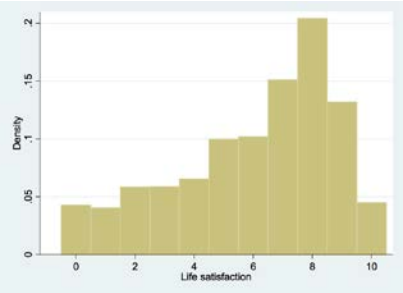
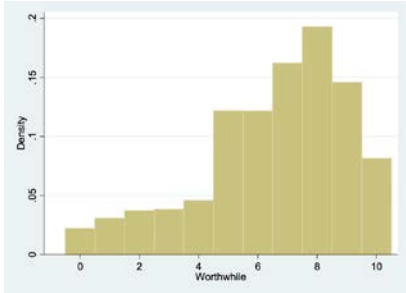
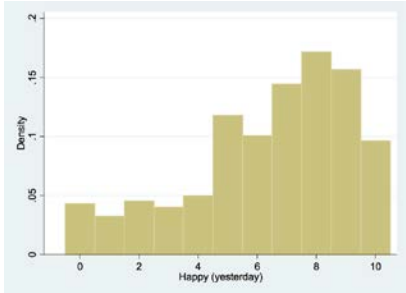
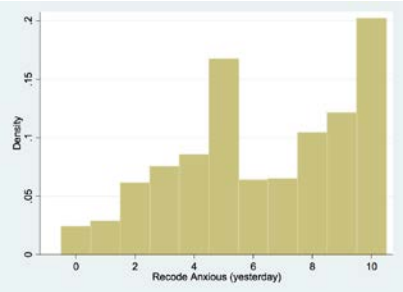
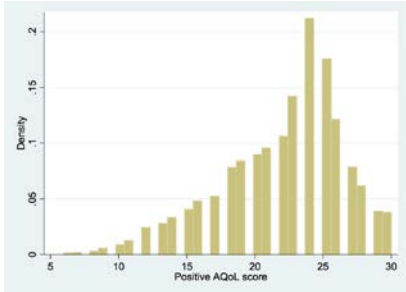
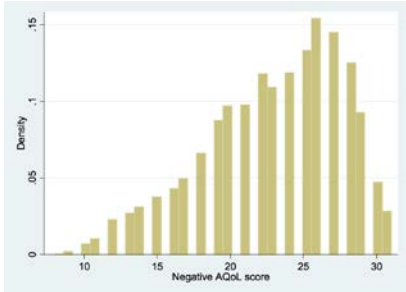
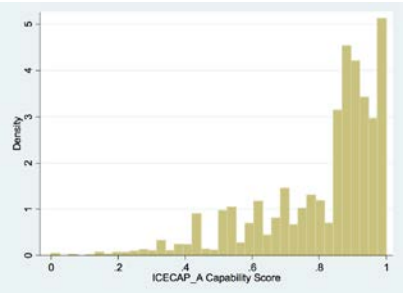
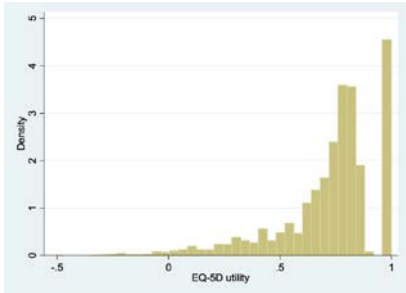
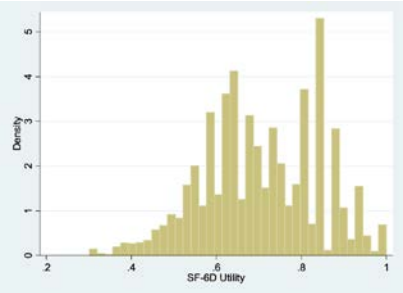
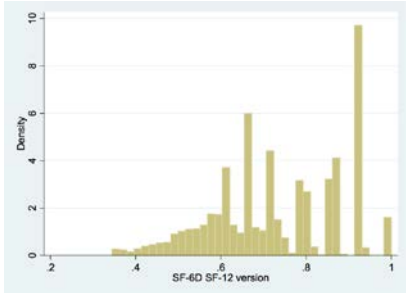
Genitourinary system	1.02	1.42	0.58	0.82	0.95	0.66	1.00
Musculoskeletal condition	0.11	0.10	0.08	0.08	0.09	0.06	1.00
Other complaints and infectious disease	1.41	2.23	0.71	0.75	0.72	0.39	1.00

ϕ GHQ scores – high scores indicate poor well-being
Figures in bold at significant at the 1% level.
Relative effect sizes calculated where EQ-5D is statistically significant.

Appendix Figure 1 Distribution of SWB and health measures - HIPO

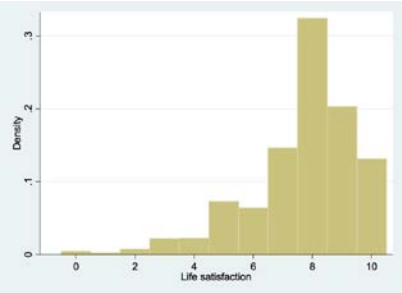
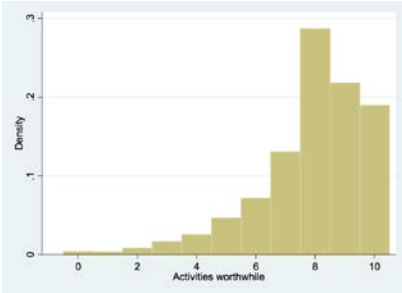
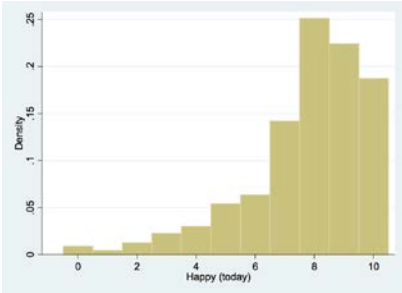
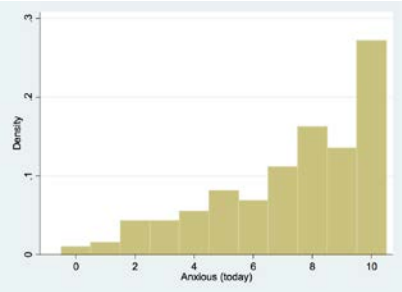
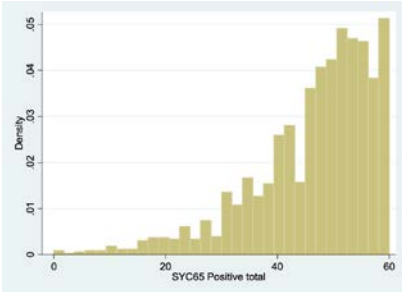
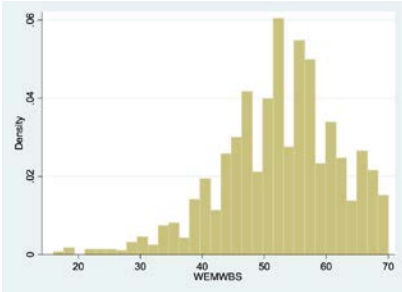
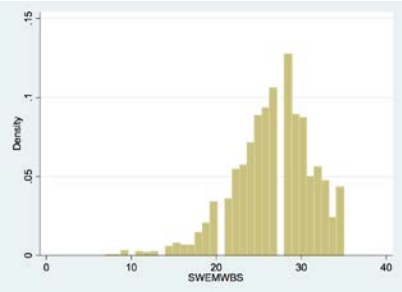
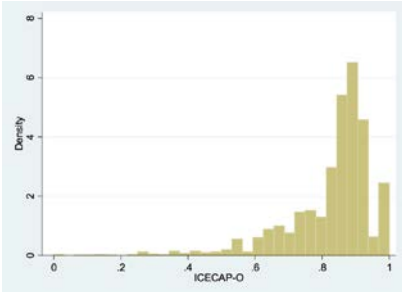
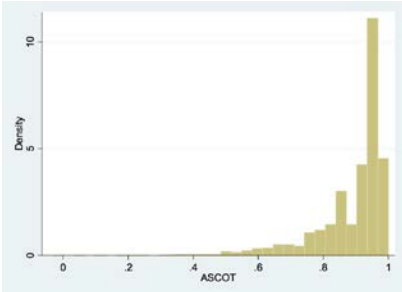
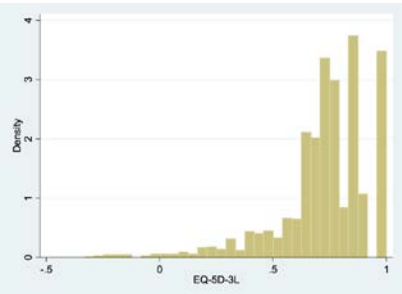
ONS-4 life satisfaction	ONS-4 – worthwhile	ONS-4 happy
		
ONS-4- anxious	Positive HIPO SWB scores	Negative HIPO SWB scores
		
EQ-5D-5L	SF-6D (SF-12)	
		
EQ-5D VAS	SWB-VAS	
		

Appendix Figure 2 Distribution of SWB and health measures – MIC

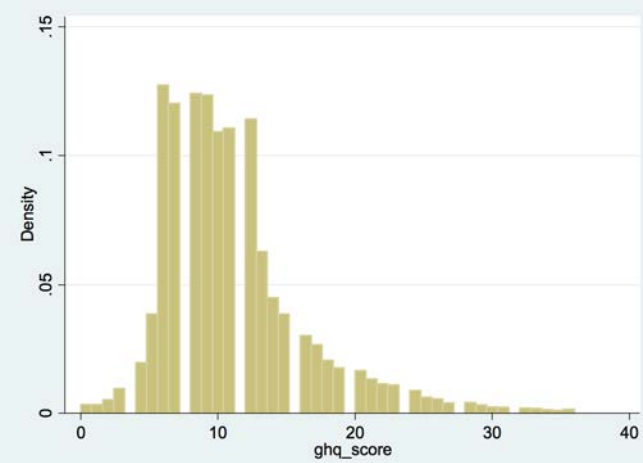
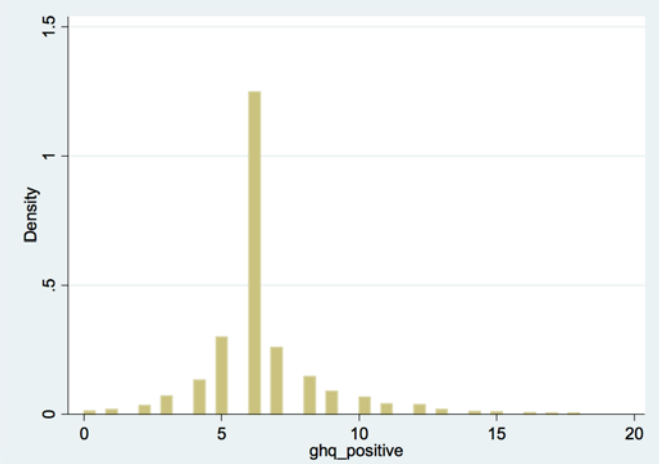
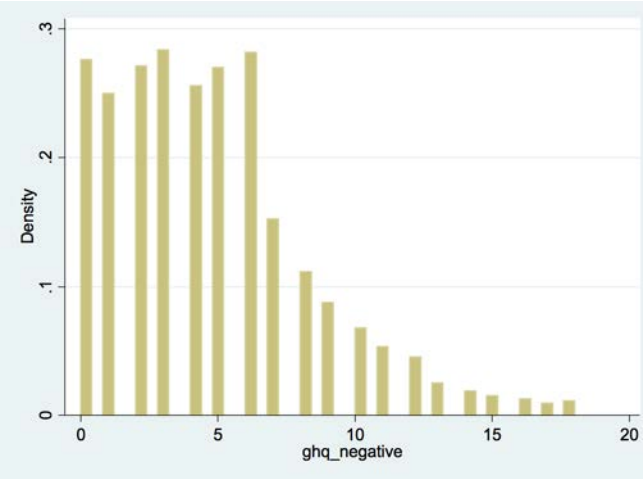
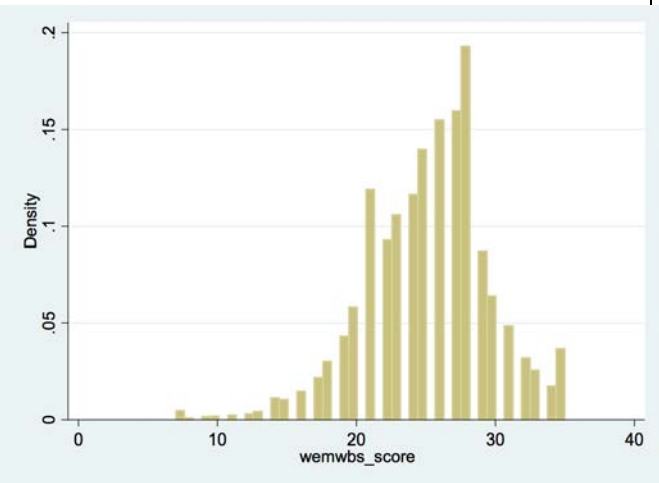
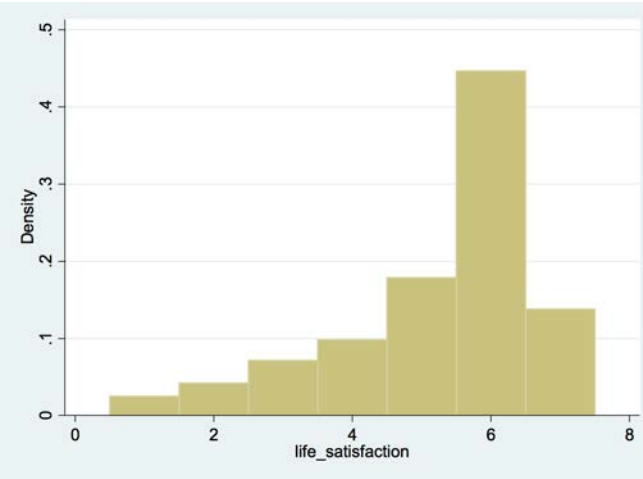
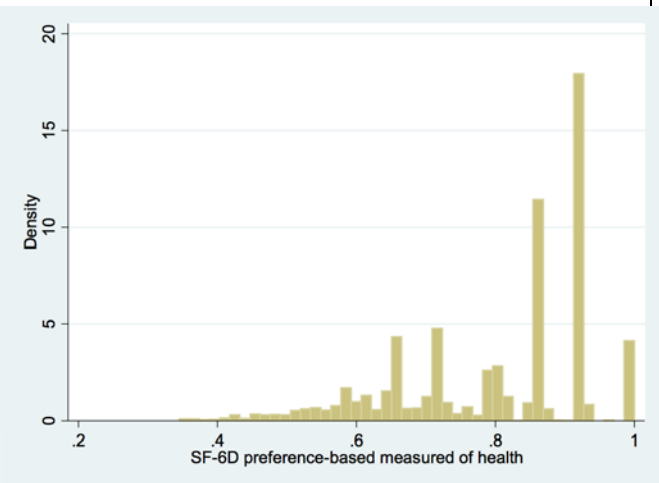
ONS-4 life satisfaction	ONS-4 – worthwhile	ONS-4 happy
		
ONS-4- anxious	AQoL positive SWB scores	AQoL negative SWB scores
		
ICECAP-A	EQ-5D-5L	
		
SF-6D (SF-36)	SF-6D (SF-12)	
		

NB: ONS-4 items in MIC had a 'neutral' label at 5

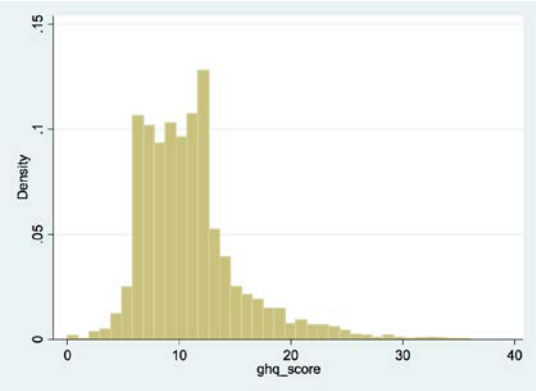
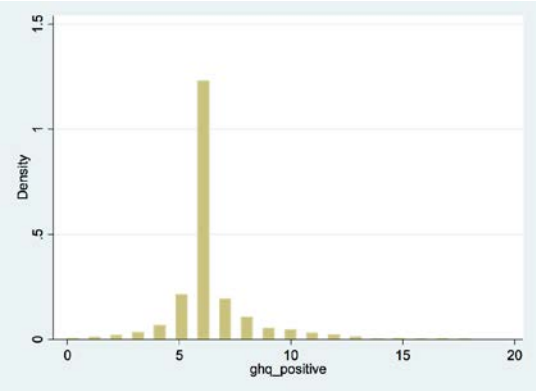
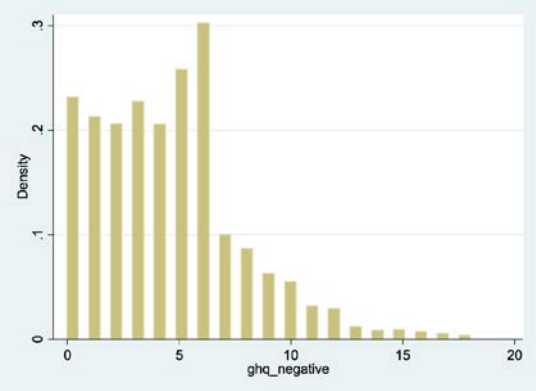
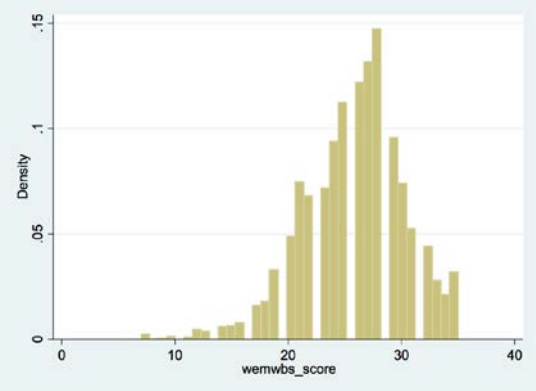
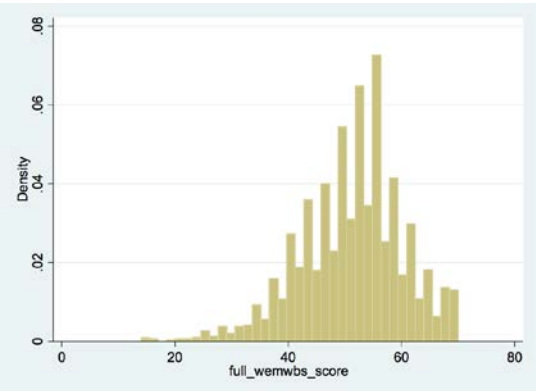
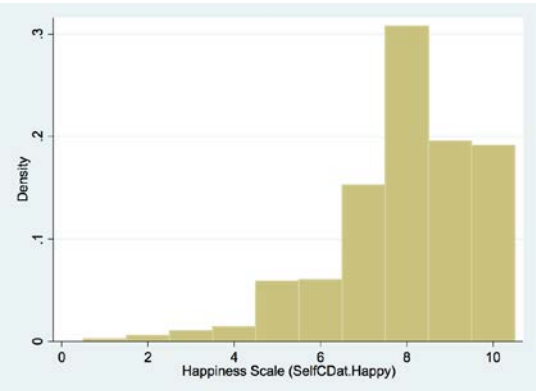
Appendix Figure 3 Distribution of SWB and health measures – SYC65

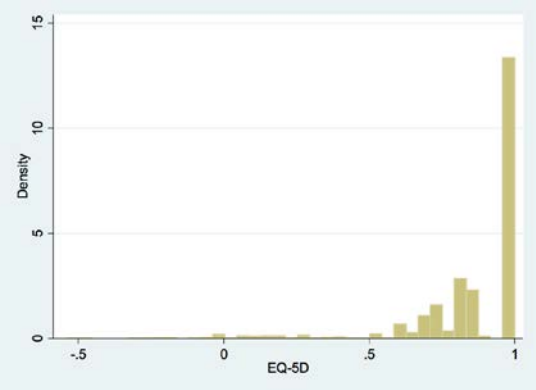
ONS-4 life satisfaction	ONS-4 – worthwhile	ONS-4 happy
		
ONS-4- anxious	SYC65 positive SWB scores	WEMWBS
		
S-WEMWBS	ICECAP-O	ASCOT
		
EQ-5D-5L		
		

Appendix Figure 4 Distribution of SWB and health measures – USoc wave 1

GHQ-12 score	GHQ positive
	
GHQ negative	**S-WEMWBS**
	
Life satisfaction (0 to 7 scale)	**SF-6D (SF-12)**
	

Appendix Figure 5 Distribution of SWB and health measures - HSE

GHQ-12 score	GHQ-positive
	
GHQ-negative	S-WEMWBS
	
WEMWBS	Happiness
	
EQ-5D-3L	


Density
0
5
10
15
-.5
0
.5
1
EQ-5D

Glossary

Term	Explanation
AQoL-8D	The Assessment of Quality of Life Instrument was developed in Australia to provide a generic measure of HRQoL suitable for economic evaluation, with the 8D version containing extensive psycho-social aspects, including subjective wellbeing.
ASCOT	The Adult Social Care Outcomes Toolkit, developed at the Personal Social Services Research Unit at the University of Kent, is a package of instruments designed to measure social-care related quality of life (SC-RQoL) and assess the extent to which an individual's social-care needs and wants are being met.
Best-worst scaling (BWS)	A type of Discrete Choice Experiment where survey respondents are shown one profile at a time and are asked to select the best and worst dimension from that profile. It is seen as less cognitively demanding that other valuation methods. The approach assumes than people's choices are a reflection of their underlying utility. Estimation of the utility model through maximum likelihood can generate weights for the dimensions and levels. The BSW approach does not anchor the score to the QALY scale.
Bifactor analysis	An extension of regular factor analysis that assumes that items load on one general dimension, with any remaining sources of covariation caused by common characteristics of items (such as measurement or nuisance factors).
Capability	Capability reflects the ability of someone to achieve a functioning rather than necessarily actually achieving that functioning. The Capability Approach, most closely linked to economist Amartya Sen, is a non-welfarist approach to social welfare analysis in which freedom to achieve well-being is of moral importance. Judgements focus on whether individuals have the opportunity to do and be what they have reason to value. The Capability Approach is closely linked to theories of justice and the measurement of poverty.
Ceiling effect	A questionnaire with ceiling effects is not able to distinguish between respondents with very high levels of the latent construct.
Confirmatory factor analysis, CFA	A type of factor analysis that adopts a clear hypothesis about the factor structure to be tested (see Factor analysis).
Convergent validity	Convergent validity refers to the extent to which the measurements from the instrument (or individual items) are related to other previously validated questionnaires or background variables related to the same construct.
Effect size	A standardised measure of the size of an effect (as opposed to the statistical significance) usually judged by the mean difference between groups divided by the standard deviation. Effect sizes provide way of distinguishing between change in the construct and measurement noise [Streiner and Norman, 2008].
EQ-5D	The EQ-5D was developed by multidisciplinary researchers in five European countries to provide an instrument with a core set of generic health status items to measure HRQoL or people's overall perceived health status.
Exploratory factor analysis, EFA	A type of factor analysis when there is no prior hypothesis about the underlying factor structure of the questionnaire (see Factor analysis).
Factor analysis	Factor analysis is a method to establish whether questionnaire items are based on a set of underlying factors or constructs.
Floor effect	A questionnaire with floor effects is not able to distinguish between respondents with very low levels of the latent construct.

HRQoL	Health-related quality of life is a multi-dimensional concept that captures the impact of health conditions and health care interventions on quality of life. This includes domains relating to physical, mental, emotional and social functioning. However, not everyone agrees that emotional and social domains should be included (Torrance 1987).
ICECAP-A	The ICECAP-A (or Investigating Choice Experiments Capability Measure for Adults) is a capability well-being measure for adults founded in Sen's capability theory.
ICECAP-O	The ICECAP-O (or Investigating Choice Experiments Capability Measure for Older people) is a capability wellbeing measure for older people founded in Sen's capability theory.
Item Response Theory (IRT)	Item response theory (IRT) is a method of modelling the responses from a questionnaire and assessing the questionnaire and each item's ability to measure underlying constructs.
Known groups	Known group validity refers to the ability of the questionnaire to detect differences between groups that are deemed to be known to be different.
ONS-4	From April 2011 the ONS included four subjective wellbeing questions on the Annual Population Survey (APS) covering an evaluative assessment of life overall, eudemonic or psychological flourishing, and positive and negative affect or day-to-day experience.
Psychometric	Psychometric analysis looks at how well a constructs (e.g. intelligence or wellbeing) is related or represented by observed response (e.g., responses on a questionnaire).
Quality Adjusted Life Year (QALY)	One QALY is equal to 1 year of life in perfect health.
Quality of life (QoL)	The terms wellbeing and quality of life are often used interchangeably as a judgement of how good an individual's life is. Quality of life is more typically used to refer to objective criteria, such as functioning and pain, whereas wellbeing can refer to objective criteria or subjective overall judgements about the individual's life.
Rasch analysis	The simplest type of IRT model for categorical data.
Reliability	Reliability refers to the consistency or repeatability of a questionnaire, meaning that similar circumstances would result in similar responses on a questionnaire.
Responsiveness	A responsive questionnaire is one that can detect changes in the relevant condition of the respondents.
Sensitivity	Sensitivity refers to the ability of a questionnaire to identify important differences in the population.
SF-36	The Medical Outcomes Study (MOS) Short Form 36-item Health Survey (SF-36) is derived from the work of the Rand Corporation during the 1970s and aimed to capture both mental and physical aspects of health.
SF-6D	The SF-6D was developed at Sheffield as a means of estimating QALY values using the most widely use health status measure, the SF-36.
Standard Gamble (SG)	A technique for measuring an individual's preferences that uses uncertainty and relies on the individual choosing between different gambles.
S-WEMWBS	A shorter 7 item version of the WEMWBS developed using Rasch analysis.
Time-trade off (TTO)	A technique for measuring an individual's preferences that does not involve uncertainty and relies on the individual choosing between different lives.
Wellbeing	The terms wellbeing and quality of life are often used interchangeably as a judgement of how good an individual's life is. We can distinguish between a number of different conceptions of wellbeing some of which focus on subjective experiences and judgements the individual makes (often referred to as subjective wellbeing), others which incorporate capabilities, objective criteria or the whether an individual gets what they want in life. During this report the use the term wellbeing in its broadest sense without aligning to any particular theory of wellbeing.

WEMWBS	The WEMWBS is a 14 item instrument developed by the Universities of Warwick and Edinburgh to identify positive mental health (i.e. identifying not just the absence of ill health but the presence of good mental health) in the general population by asking questions in a positive manner.