EEPRU

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

Research
Interim Report

An empirical comparison of well-being measures used in UK

Clara Mukuria, Donna Rowen, Tessa Peasgood, John Brazier

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

RR 0027

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 these six commonly used measures of health and well-being: SWEMWBS, GHQ-12, ONS-4, ICECAP-A, 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 SWEMWBS are measured. GHQ-12 has both negative and positive items while the SWEMWBS 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 practioners 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, SWEMWBS, ICECAP-A/O) and health (EQ-5D and SF-6D) and social care (ASCOT) measures and looking at correlations between the 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 SWEMWBS 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 SWB measures were strongly correlated, particularly when considering positive SWB items and measures (ONS-4 life satisfaction, happiness, worthwhile, ICECAP-A/O). The relationship was moderate for the 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.

Factor analysis suggested that there were one or two SWB factors related to positive items with a separate factor for negative items. However, this was not conclusive across datasets and may have been influenced by the response options of different measures. For example, factor analysis of the GHQ-12 and SWEMWBS resulted in two positive SWB factors linking to positive items from the GHQ and SWEMWBS 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. Based on these results it is difficult to judge whether or not the ONS should include GHQ-12 alongside the SWEMWBS. 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. This means that the negative SWB questions may reveal something beyond positive SWB, suggesting limitations in the use of SWEMWBS alone which contains only positive items. However, there are conceptual and empirical concerns with the GHQ-12 and so it would not be recommended for this to be used alone either.

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, with GHQ-12 (and GHQ negative) doing better than the health measures, as did some of the aggregate positive and negative SWB scores: ICECAP-O/A did better than EQ-5D-5L, SWEMWBS 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. This finding was confirmed when reported disability was used with SWB measures better able to reflect memory or concentration problems than the health measures while the reverse was true for disability with physical health. Panel data confirmed 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

Although the analysis undertaken focused on measures that are currently used or recommended for use in the UK using large patient and general population datasets, 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 SWEMWBS 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 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, SWEMWBS 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 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.

Future work will be undertaken to address some of the limitations, such as undertaking confirmatory factor analysis which will include correlation of errors across items that are correlated as this may affect results. Methods factors will also be tested to assess whether or not the way questions are asked has an impact on the factor analysis results. We will also assess the use of an aggregate ONS-4 positive item, excluding the anxiety question which performs differently.

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 (SWEMWBS) and the General Health Questionnaire (GHQ-12), which is a mental health screening measure that has been used in well-being measurement. 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, 2012; 2013a] while social care guidance 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]

There is limited evidence on how the GHQ-12, ONS-4, SWEMWBS, 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 SWEMWBS is included alongside the ONS-4 within the ONS well-being indicator set.

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 these six commonly used measures of health and well-being: SWEMWBS, GHQ-12, ONS-4, ICECAP, ASCOT and EQ-5D. This report addresses the second task by providing an interim 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 SWEMWBS are measured. GHQ-12 has both negative and positive items while the 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. 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. A further report will present some more advanced psychometric analyses.

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 SWEMWBS) 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 (SCRQoL). 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. It covered most specialties, though patients with a primary mental health diagnosis are not included in the survey. 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 6,452 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. Data cleaning resulted in exclusion of 101 respondents who said

they were completely happy/satisfied/life was going well (score 10) as well as completely anxious (score 10), or that they were completely bored as well as completely enjoying their activities (score 10). Respondents with inconsistent responses had lower EQ-5D scores (0.436 vs. 0.673, t_{6236} = -8.0, p<0.001). Of the remaining 6,302, 1,519 (24%) had missing data in one or more of the questions asked, and they were older (64.0 vs. 59.2, t_{6300} =-10.1, p<0.001), less healthy (EQ-5D-5L: 0.596 vs. 0.695, t_{6094} = 11.2, p<0.001) and had lower SWB (life satisfaction: 6.13 vs. 6.78, t_{6195} = 8.0, p<0.001; anxious: 6.97 vs. 7.58, t_{6197} = 6.9, 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.

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), SF-36, self-complete time trade off 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. There will be a follow-up at 12 months to provide longitudinal information on changes in health and social care. 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) but had only slightly 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 (Appendix Table 1).

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, SWEMWBS, 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.

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

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 when you are in physical danger, your 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 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, 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 (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 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. 47.9, 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. This 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 identify those in need of psychiatric care.[Goldberg and Williams, 1998] 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 takes 'same as usual' for the negative items as an indication of the

presence of a symptoms.[Goodchild and Duncan-Jones, 1985] Likert scoring was reported and caseness scoring was analysed but there were no differences in the results so these are not reported (results available from the authors). We scored the six positive and six negative items separately using Likert scoring purely for data exploration purposes. 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 SWEMWBS

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. A shortened 7-item version, the SWEMWBS, 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] A simple aggregate scoring was used for this 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] The measures are scored on a range from 0 to 1, where 0 represents no capability and 1 represents full capability.[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 by summing across the items, as in 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 of 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 from the general population that range from -0.594 to 1,[Dolan, 1997] and a cross-walk algorithm can be used to generate 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. The EQ-5D covers

depression/anxiety in its descriptive part as well as elements of preference satisfaction in its scoring.[Peasgood et al, 2014] 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. 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 had SF-6D based on the SF-12, while MIC had 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 SCRQoL that is designed to assess the extent to which an individual's social-care needs and wants are being met. It also draws upon Sen's theory of capabilities, with the level of met need in a domain being disaggregated into having needs met and an 'ideal' state in which the level of functioning in the domain is consistent with individual preferences. This suggests that while capabilities matter, the individual with low levels of functioning can be judged by society as having an unacceptable level of need regardless of whether they recognize this to be the case.[Netten et al, 2011] It 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). For each item the level of met need is assessed across four levels: ideal, no unmet needs, some unmet needs and high unmet needs (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). 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 BWS; another has anchored 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.[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 construct(s) given their different theoretical basis; 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, provide a measure of potential similarities. High floor/ceiling effects (when over 5% of respondents are at the floor or ceiling of a measure[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

[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.

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 self-report biases (such as social desirability, avoiding the ends of scales, 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 are due to the error rather than 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). However, the fact that the well-being instruments have different distributions needed to be addressed. Difference based effect sizes (such as Cohen's d) allow standardised comparisons of means to be made which take into account the distributions. However, these 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] Differences in SWB for groups with a health condition may therefore be affected by age which may bias the results. Cohen's d effect sizes[3] for self-reported health conditions are reported for USoc data. These were based on matched data (one-to-one matched using three age categories and gender). These are reported in

[3] 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.

order to demonstrate the magnitude of the effect sizes in the most familiar metric, Cohen's d, in which 0.2 is usually taken to be a small effect, 0.5 considered medium and 0.8 large. [Cohen, 1992]

However, an easy way to more thoroughly 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. Eta^2 was used as the measure of effect size: the proportion of the total variation in the dependent variable that can be attributed to an independent variable (ranging from 0 to 1)[4]. The interest here is in making comparisons between the effect sizes (Eta^2) within the same dataset. Although there are particular rules of thumb for assessing Eta^2, this is not the focus of the analysis as we are interested in relative performance of SWB measures to health measures. For example, the question being asked is whether the presence of a particular health condition explains more variation in EQ-5D than it explains variation in life satisfaction. To address this, the SWB effect sizes were compared to the EQ-5D (or SF-6D if EQ-5D was not available) effect sizes using relative effect sizes (SWB eta squared/EQ-5D or SF-6D Eta^2). This allows us to judge the relative impact of a move from health to SWB measures. Ordinary least squares regressions were undertaken with Eta^2 effect sizes calculated 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.

Analysis based on cross-section data may suffer from bias arising from unobserved individual effects. Individuals with particular personality 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 2 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.

[4] Eta^2 = Sum of Squares of the effect / Sum of Squares of the total. In small samples, eta squared is an upwardly biased estimation of strength of association but it was used in this study to describe differences within large samples.

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]. Again the focus here is on the relative size of the Cohen's f^2 compared to that for the SF-6D. 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 Exploratory factor analysis

Exploratory factor analysis was used to explore the dimensional structure of the measures to identify whether different SWB items were measuring the same underlying concept within a measure, across different measures and by positive and negative wording. Items for the two health measures (EQ-5D and SF-12) and ASCOT were also included in the analysis. Factor analysis aims to identify underlying unobservable (latent) variables or domains, that is, factors that cause the observed variables 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. The recommended cut offs for considering when an item loads onto a factor vary in the literature. We use a commonly accepted one with 0.8 taken as high, and 0.32 as a minimum. Where an item loads at 0.32 or higher onto two or more factors, the item is considered to "cross load" and may not be associated with a single factor.[Tabachnick and Fidell, 2001] The level of variance in each item that is not explained by the factors, referred to as uniqueness, is also reported. A uniqueness level of more than 0.6 is considered high and was used to flag up items that may be measuring something outside the common factors. Factors were assumed to be correlated as it is unlikely that factors emerging from measures of well-being or health would be independent of each other. Correlation between factors is reported; strong correlations (rho≥0.5) may indicate that factors are tapping into the same latent construct.

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

[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 direct comparisons that can be performed. HIPO, SYC65 and MIC allowed comparisons of ONS-4[6] and the health measures; USoc and HSE allowed comparison of GHQ-12 and SWEMWBS 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	SWEMWBS	GHQ-12	ICECAP-A/O	ASCOT	EQ-5D
SWEMWBS	x					
GHQ-12	x	USoc[7], HSE				
ICECAP-A/O	MIC, SYC65	SYC65	x			
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 are used to provide information for all 3 questions. Strong correlations indicate similarities in measures. Factor analysis provides 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 assess differences in the same way as health measures. Relative effect sizes greater than 1 indicate that SWB are more sensitive, while less than 1 indicates that they are less sensitive than the health measures.

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?			√	√	√

[6] 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.

[7] A separate analysis was also undertaken for the BHPS wave 18 which has similar SWB measures to the USoc but it did not include any health measure nor did it add anything different to the results so this has not been reported.

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).

There was evidence of ceiling effects (i.e. large proportion at the best score indicating high levels of SWB) in the single life satisfaction, happiness, worthwhileness and anxiety questions in HIPO, SYC65, but fewer for 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.

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

	mean	SD	min	max	median	25 per.	75 per.	% at floor	% at ceiling	% missing
HIPO n=5,344										n=6,351
Well-being										
ONS-4										
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
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
Health										
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
VAS										
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										
Well-being										

ONS-4										
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
Positive AQoL										
SWB total	22.04	4.54	6	30	23	19	25	0.00	2.39	NA
Negative AQoL										
SWB total	22.95	4.73	8	31	24	20	27	0.00	1.72	NA
ICECAP-A	0.81	0.18	-0.001$^{\phi}$	1	0.88	0.71	0.95	0.13	11.99	NA
Health										
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	0.15	1.26	NA
SF-6D (SF-36)	0.70	0.14	0.301	1	0.70	0.61	0.81	0.00	2.78	NA

Table 3 (continued): Well-being and health measures summary statistics (SYC65)

	mean	SD	min	max	median	25th per.	75th per.	% at floor	% at ceiling	% missing
SYC65 n=1593										n=1749
Well-being										
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
Positive SYC65										
SWB total	46.70	10.81	0	60	49	42	55	0.06	7.22	1.32
SWEMWBS	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
Health										
EQ-5D-5L	0.73	0.21	-0.33	1	0.76	0.66	0.84	0.00	14.50	1.14
VAS										
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

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 responses of 'same as usual' for positive items (USoc: 47%; HSE: 58%) and 'not at all' or 'no more than usual' for negative items (USoc: 61%; HSE: 69%). SWEMWBS 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 not comparable to the level of missing data for health. However, the health questions were asked in the interview whereas the SWB questions were self-complete which makes it difficult to compare them. WEMWBS and the happiness question also had a high level of missing data.

GHQ-12 overall scores and WEMWBS had lower proportions at the top end of the scale but GHQ negative had higher 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										
Well-being										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
SWEMWBS	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
Health										
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
Well-being										
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
Health										
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.

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

3.2 Correlations

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 were 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 SWEMWBS 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 SWEMWBS and the GHQ-positive sub-score compared to SWEMWBS 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 SWEMWBS 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 about frequency, but the response options of the GHQ-12 means that the relationship between the two measures is weaker than expected.

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 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 was 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 measuring the same thing.

Table 5: Spearman correlations between well-being items (ONS 4, individual well-being items), health (EQ-5D-5L, SF-6D) and social care (ASCOT) measures (HIPO, MIC and SYC65)

HIPO n =5,344	satisfaction	worthwhile	happy	recoded anxious	ONS-4 total	Positive HIPO SWB total	Negative HIPO SWB total
Well-being							
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		
Positive total	0.84	0.80	0.85	0.64	0.88	1.00	
Negative total	0.67	0.64	0.70	0.70	0.77	0.72	1.00
Health							
EQ-5D-5L	0.63	0.52	0.56	0.47	0.62	0.63	0.55
SF-6D (SF-12)	0.72	0.62	0.68	0.58	0.74	0.73	0.67
EQ-VAS	0.70	0.58	0.63	0.48	0.67	0.67	0.57
SWB-VAS	0.82	0.73	0.79	0.61	0.83	0.84	0.68

MIC n =6,808	satisfaction	worthwhile	happy	recoded anxious	ONS-4 total	Positive AQoL SWB total	Negative AQoL SWB total	ICECAP-A
Well-being								
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			
Positive total	0.70	0.65	0.67	0.34	0.72	1.00		
Negative total	0.62	0.54	0.62	0.42	0.69	0.81	1.00	
ICECAP-A	0.65	0.58	0.60	0.31	0.65	0.81	0.74	1.00
Health								
EQ-5D-5L	0.39	0.31	0.35	0.24	0.40	0.54	0.54	0.56
SF-6D (SF-12)	0.51	0.45	0.50	0.39	0.58	0.66	0.70	0.66
SF-6D (SF-36)	0.48	0.41	0.46	0.36	0.54	0.63	0.66	0.64

SYC65 n=1593	satisfaction	worthwhile	happy	recoded anxious	ONS-4 total	Positive SWB total	SWEMWBS	WEMWBS	ICECAP-O
Well-being									
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				
Positive total	0.79	0.76	0.77	0.50	0.79	1.00			
SWEMWBs	0.66	0.68	0.69	0.52	0.73	0.78	1.00		
WEMWBS	0.68	0.69	0.71	0.51	0.74	0.80	0.96	1.00	
ICECAP-O	0.63	0.59	0.60	0.47	0.65	0.71	0.68	0.70	1.00
Health/social care									
EQ-5D-5L	0.44	0.36	0.40	0.37	0.47	0.49	0.43	0.46	0.47
ASCOT	0.48	0.44	0.45	0.34	0.49	0.54	0.49	0.51	0.56
VAS									
EQ-VAS	0.59	0.50	0.54	0.41	0.59	0.64	0.55	0.59	0.57
WB-VAS	0.74	0.65	0.70	0.50	0.74	0.76	0.68	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	SWEMWBS score	Life satisfaction
Well-being					
GHQ score ϕ	1				
GHQ positive ϕ	0.86	1			
GHQ negative ϕ	0.95	0.65	1		
SWEMWBS score	-0.61	-0.50	-0.59	1	
Life satisfaction	-0.49	-0.40	-0.48	0.50	1
Health					
SF-6D (SF-12)	-0.56	-0.47	-0.54	0.42	0.36

HSE n = 5,709	GHQ score	GHQ positive	GHQ negative	SWEMWBS score	WEMWBS score	Happy
Well-being						
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
Health						
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

3.3 Factor analysis

Factor analysis was used to assess whether moderate correlations indicated that measures cover different dimensions. The number of factors identified depended on which dataset was used.

There were 3 factors in HIPO. There was one factor with physical health items from the SF-12 and EQ-5D including the social functioning question from the SF-12 (Table 7). This factor also included items on 'feeling capable' as well as 'feeling tired', though there was evidence of cross-loading for the latter. There were two SWB factors: one factor had the negative items on anxiety, loneliness, anger, emotional health as well as an item on feeling calm; the other had evaluative aspects of SWB including life satisfaction and enjoying activities (Table 7). This third factor also had items related to social functioning from the single item SWB questions. However, many items cross-loaded on more than one factor, so a clear factor structure was not identified. Life satisfaction and worthwhile did not load fully on either of the SWB factors. As with correlation analysis, the two SWB related factors had slightly larger correlations (ρ=0.64) compared to correlations between these factors and the health factor (ρ= 0.58 and 0.50 respectively).

Table 7: Rotated component matrix for factor analysis on well-being items (HIPO)

Variable	Factor1	Factor2	Factor3	Uniqueness	
n=5,158					
Capable	0.6046		0.3019	0.3184	
rsf1 General health	0.6667			0.3441	
sf2 Limited moderate physical	0.9554			0.2292	
sf3 Limited strenuous physical	0.9178			0.2969	
sf4 Accomplished less physical	0.8902			0.1932	
sf5 Limited work physical	0.9179			0.1777	
rsf8 Pain interfered work	0.8063			0.2916	
rtired	0.3600	0.3579		0.6264	*cross-loader*
rsfEnergy	0.5740			0.3995	
sf12 Social activities	0.5627	0.3375		0.2858	*cross-loader*
req Mobility	0.9130			0.2608	
req Self-care	0.6900			0.4698	
req Usual activities	0.8915			0.2176	
req Pain/discomfort	0.7795			0.3526	
ONS-4 happy		0.5605	0.4013	0.1642	*cross-loader*
content		0.5555	0.4080	0.1626	*cross-loader*
rlonely		0.6943		0.5123	
rangry		0.8469		0.4420	
rbored		0.6068		0.4959	
ONS-4 ranxious		0.9660		0.2505	
sf6 Accomplished less emotional	0.3973	0.6492		0.2918	*cross-loader*
sf7 Less careful work emotional	0.4149	0.6224		0.3149	*cross-loader*
rsfCalm		0.6685		0.4342	
sf Downhearted		0.7849		0.4327	
req Anxiety/Depression		0.7845		0.3296	
look forward to future		0.3506	0.6007	0.2678	*cross-loader*
have social support			0.6921	0.4033	
contribute to others happiness			0.8966	0.4675	
enjoy activities			0.5758	0.2751	
life going well		0.3178	0.5529	0.1822	*cross-loader*
ONS-4 life satisfaction		0.3707	0.4268	0.2021	*cross-loader*
ONS-4 worthwhile		0.3930	0.4888	0.2487	*cross-loader*

Correlations	Factor 1	Factor 2	Factor 3
Factor 1	1		
Factor 2	0.581	1	
Factor 3	0.496	0.636	1

Factor loading |<0.3| are not reported; Extraction Method: Principal Component Factors. Rotation Method: Promax with Kaiser Normalization. Uniqueness >0.6 indicates item may not be contributing to the factor
Cross-loading: factor loading on two or more factors > 0.32
Prefix: sf(rsf) – recoded SF-12 items (recoded) req – recoded EQ-5D items

In MIC, there were four factors. The first two were distinct SWB factors: one with the positive SWB items from the ONS-4 and AQoL along with isolation from the AQoL; and one with negative SWB items from AQoL and some negative SWB from the health measures (Table 8). The third factor was a health factor with SF-12 general health and physical functioning, four EQ-5D dimensions and items related to 'energy' and 'independence' which did not load as strongly on the health factor (Table 8). The final factor had emotional functioning questions from the SF-12 alongside pain and social activities which may have been due to the focus on role limitations or interference with work that was common in the wording of these questions. There was some evidence of cross-loading and the 'independence' question did not fit into the 3rd factor as well as the other items (uniqueness>0.6). The first two SWB factors were strongly correlated (ρ=0.66) while correlations with the other factors were moderate (ρ=0.39 to 0.44).

Table 8: Rotated component matrix for factor analysis on well-being and health items (MIC)

Variable	Factor1	Factor2	Factor3	Factor4	Uniqueness	
n=6808						
ONS-4 life satisfaction	0.7142				0.3539	
ONS-4 worthwhile	0.7787				0.3989	
ONS-4 happy	0.5822				0.4053	
raq Enthusiastic	0.5743				0.3746	
raq Happy	0.6600				0.2783	
raq Enjoy Relationships	0.9196				0.3934	
raq Pleasure	0.7028				0.3869	
raq Control	0.5142	0.3315			0.3487	*cross-loader*
raq Content	0.6971				0.2152	
ric Settled	0.5489				0.3711	
ric Friendship	0.9502				0.3550	
ric Achievement	0.6035				0.4079	
ric Enjoyment	0.7679				0.3243	
raq Isolated	0.5605				0.4026	
ONS-4 anxious recoded		0.7856			**0.5591**	
raq Sad	0.3422	0.5401			0.2958	*cross-loader*
raq Worried		0.7254			0.3091	
raq Depressed		0.5399			**0.5594**	
raq Despair		0.5228			0.3094	
raq Angry		0.4806			0.6368	
rsf Calm and peaceful		0.5846			0.4169	
sf Downhearted		0.5441			0.3069	
req Anxious/Depressed		0.5735			0.2772	
rsf1 General health			0.6533		0.3783	
sf2 Limited moderate physical			0.8652		0.2951	
sf3 Limited strenuous physical			0.8577		0.3500	
sf4 Accomplished less physical			0.5559	0.4768	0.2915	*cross-loader*
sf5 Limited work physical			0.6539	0.4491	0.2485	*cross-loader*
rsf Energy		0.3351	0.4411		0.4355	*cross-loader*
req Mobility			0.9082		0.2572	
req Selfcare			0.5991		0.5643	
req Usual Activities			0.8064		0.2573	
req Pain/Discomfort			0.7835		0.3913	
ric Independent			0.3946		**0.6273**	
sf6 Accomplished less emotional				0.7657	0.2100	
sf7 Less careful work emotional				0.8073	0.2076	
rsf8 Pain interfered work				0.4984	0.3398	
sf12 Social activities				0.4882	0.3435	
Factor 1	1					
Factor 2	0.661	1				
Factor 3	0.424	0.390	1			
Factor 4	0.438	0.434	0.437	1		

Factor loading |<0.3| are not reported; Extraction Method: Principal Component Factors. Rotation Method: Promax with Kaiser Normalization. Uniqueness >0.6 indicates item may not be contributing to the factor

Cross-loading: factor loading on two or more factors > 0.32

Prefix: raq – recoded AQoL items from the happiness and mental health dimensions; ric – recoded ICECAP-A items
sf(rsf) –SF-12 items (recoded); req – recoded EQ-5D items

SYC65 had six factors with three related to SWB: as with HIPO and MIC, there was a factor that was related to life satisfaction and life going well, a factor related to items from the WEMWBS which was around thinking and dealing with problems; and a final factor related to anxiety and lack of security (Table 9). There were a number of items related to feeling optimistic, enjoyment, control and security mainly from the ICECAP-O which did not load strongly onto any factor (Table 9). There was also evidence of cross-loading. A fourth factor was related to social functioning. The EQ-5D items apart from anxiety/depression, physical health item and energy loaded on a separate factor in SYC65 (Table 9). This factor also had personal safety and the ICECAP-O item on role but these did not load strongly on any factor. There was also a factor related to accommodation, food and cleanliness. There was some evidence of cross-loading. Correlations were strong between the first two SWB factors (ρ = 0.66), while they were small to borderline strong for the other factors (ρ= 0.1 to 0.5, Table 9).

In USoc there were four SWB factors: one with the GHQ-negative items along with the SF-12 item 'calm and peaceful', one which combined life satisfaction with the SWEMWBS, one which had the SF-12 items in emotional health and social functioning, and one related to the GHQ positive items which also contained the negative item 'feeling worthless' (Table 10). The final factor was a physical health factor based on the SF-12 physical health items. Correlations were strongest between the first factor with GHQ negative items and the next two factors with WEMWBS and SF-12 emotional health items (ρ=-0.52 and -0.53 respectively). The other correlations were small to moderate (ρ= |0.23 to 0.49| (Table 10). The factor analysis results for HSE were similar to USoc with three SWB factors related to negative GHQ, including the happiness item, one factor related to the SWEMWBS items and a factor related to the positive GHQ items. There was a fourth factor related to the physical health items from the EQ-5D (Table 11). There was less cross-loading in USoc and HSE compared to HIPO and SYC65. Correlations were strong for the GHQ negative and SWEMWBS factors (ρ=0.55) and ranged from small to borderline strong for the other factors (ρ=|0.24 to 0.50|) which is similar to the USoc results.

Table 9: Rotated component matrix for factor analysis on well-being and health items (SYC65)

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Uniqueness	
n=1596								
Life satisfaction	0.8148						0.1946	
Activities worthwhile	0.8138						0.2425	
Happy (today)	0.7609						0.1930	
Can do things	0.5997		0.3576				0.2888	*cross-loader*
Looking forward to tomorrow	0.8177						0.2538	
Supportive relationships	0.6169					0.3652	0.2871	*cross-loader*
Contribute to others happiness	0.7077						0.3155	
Enjoy activities I do	0.8058						0.2135	
Life is going well	0.7557						0.1480	
Wbs Optimistic	0.3401	0.3398					0.5476	*cross-loader*
ICap Enjoyment	0.3357						0.4345	
Wbs Useful	0.3608	0.4535					0.3541	*cross-loader*
Wbs Relaxed		0.5018		0.3347			0.3628	*cross-loader*
Wbs Interested in others		0.5375		-0.3147		0.3464	0.4586	*cross-loader*
Wbs Dealing with problems well		0.7043					0.3434	
Wbs Thinking clearly		0.8240					0.3113	
Wbs Feeling good about self		0.5767					0.2712	
Wbs Feeling confident		0.6584					0.2598	
Wbs Able to decide		0.7999					0.3356	
Wbs Interested new things		0.6445					0.4441	
Wbs Feeling cheerful		0.4923					0.2955	
ICap Control		0.3570	0.3527				0.4771	*cross-loader*
req Mobility			0.9330				0.2294	
req Self-care			0.5405		0.3799		0.4035	
req Usual Activities			0.8253				0.2060	
req Pain Discomfort			0.8207				0.3687	
Physical health interferes			0.8634				0.2897	
Wbs Energy		0.4006	0.5612				0.3745	*cross-loader*
rAsc Personal safety			0.3989		0.3031		0.5369	*cross-loader*
ICap Role	0.2190	0.2162	0.2894				0.4674	
req Anxiety Depression				0.6273			0.3279	
Mental health interferes				0.5775			0.4397	
rAnxious (today)				0.7489			0.4041	
ICap Security				0.4115			0.5483	
rAsc Control					0.5256		0.4205	
rAsc Personal cleanliness					0.7484		0.3598	
rAsc Food & drink					0.7498		0.4667	
rAsc Occupation			0.2732		0.3021		0.4743	*cross-loader*
rAsc Accommodation					0.6326		**0.5617**	
Wbs Feeling close to others		0.4505				0.5334	0.2874	*cross-loader*
Wbs Feeling loved						0.7326	0.2828	
ICap Love friendship						0.7401	0.3130	
rAsc Social participation					0.4047	0.4071	0.4511	*cross-loader*
rAsc Dignity					0.3389	0.3475	**0.8109**	*cross-loader*

Factor 1	1					
Factor 2	0.656	1				
Factor 3	0.503	0.426	1			
Factor 4	0.500	0.423	0.362	1		
Factor 5	0.432	0.339	0.452	0.398	1	
Factor 6	0.462	0.393	0.123	0.208	0.104	1

Factor loading |<0.3| are not reported; Extraction Method: Principal Component Factors. Rotation Method: Promax with Kaiser Normalization. Uniqueness >0.6 indicates item may not be contributing to the factor
Cross-loading: factor loading on two or more factors > 0.32

Prefix: wbs –WEMWBS items; req – EQ-5D items (recoded)

Table 10: Rotated factor loadings for factor analysis on well-being items (USoc)

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Uniqueness	
n=37554							
ghq2 lost sleep over worry	0.7612					0.4287	
ghq 5 constant strain	0.8968					0.3281	
ghq 6 couldn't overcome difficulties	0.6163					0.4585	
ghq 9 unhappy/depressed	0.6401					0.3082	
ghq 10 losing confidence	0.4553					0.4113	
swemwbs 3 feeling relaxed	-0.5226	0.5155				0.3617	*cross-loader*
rsf Calm and peaceful	-0.5985					0.5495	
swemwbs 1 optimistic about future		0.6575				0.5702	
swemwbs2 feeling useful		0.7496				0.4104	
swemwbs 4 dealing with problems		0.7140				0.3513	
swemwbs 5 thinking clearly		0.7244				0.3395	
swemwbs 6 close to others		0.7747				0.4626	
swemwbs 7 make up my own mind		0.7219				0.4529	
Life satisfaction		0.3591				**0.6211**	
sf6 Accomplished less emotional			0.8662			0.1763	
sf7 Less careful work emotional			0.8938			0.1908	
sf Downhearted	-0.3231		0.6136			0.3263	*cross-loader*
sf12 Social activities			0.5211		0.3676	0.3694	*cross-loader*
ghq 11 feeling worthless				0.3128		0.4788	
ghq1 concentrate				0.5212		0.5409	
ghq 3 play useful part				0.7142		0.4901	
ghq 4 capable decisions				0.7814		0.4437	
ghq 7 enjoy activities	0.3600			0.4994		0.4708	*cross-loader*
ghq 8 able to face problems				0.7179		0.4438	
ghq 12 happy				0.5441		0.4864	
rsf1 General health					0.6741	0.4734	
sf2 Limited moderate physical					0.8408	0.3133	
sf3 Limited strenuous physical					0.8417	0.3245	
sf4 Accomplished less physical					0.8257	0.2532	
sf5 Limited work physical					0.8506	0.2062	
rsf8 Pain interfered work					0.7211	0.4653	
rsf Energy					0.5920	0.5016	
Factor 1	1						
Factor 2	-0.515	1					
Factor 3	-0.534	0.408	1				
Factor 4	0.491	-0.407	-0.442	1			
Factor 5	-0.257	0.227	0.433	-0.308	1		

Factor loading |<0.3| are not reported; Extraction Method: Principal Component Factors. Rotation Method: Promax with Kaiser Normalization. Uniqueness >0.6 indicates item may not be contributing to the factor

Cross-loading: factor loading on two or more factors > 0.32

Prefix: ghq – GHQ12 items; swemwbs – Short WEMWBS items; sf(rsf) –SF-12 items (recoded)

Table 11: Rotated factor loadings for factor analysis on well-being items (HSE)

Variable	Factor1	Factor2	Factor3	Factor4	Uniqueness	
n=5709						
ghq2 lost sleep over worry	0.8415				0.4383	
ghq 5 constant strain	0.8710				0.3739	
ghq 6 couldn't overcome difficulties	0.6936				0.4104	
ghq 9 unhappy/depressed	0.7894				0.2917	
ghq 10 losing confidence	0.6096				0.3948	
ghq 11 feeling worthless	0.5211				0.459	
req Anxiety/Depression	-0.5366				0.5091	
Happiness	-0.4448	0.3124			0.5073	*cross-loader*
swemwbs 1 optimistic about future		0.6812			0.5542	
swemwbs2 feeling useful		0.7788			0.4119	
swemwbs 3 feeling relaxed	-0.3755	0.5329			0.4500	*cross-loader*
swemwbs 4 dealing with problems		0.7354			0.3456	
swemwbs 5 thinking clearly		0.7232			0.3707	
swemwbs 6 close to others		0.7496			0.4669	
swemwbs 7 make up my own mind		0.7197			0.4333	
ghq1 concentrate			0.5854		0.5473	
ghq 3 play useful part			0.7575		0.4383	
ghq 4 capable decisions			0.8455		0.3756	
ghq 7 enjoy activities			0.541		0.4799	
ghq 8 able to face problems			0.6564		0.4678	
ghq 12 happy	0.3200		0.4968		0.5010	*cross-loader*
req Mobility				0.8598	0.2843	
req Self-care				0.6753	0.5127	
req Usual activities				0.8343	0.2634	
req Pain/Discomfort				0.7891	0.3783	
Factor 1	1					
Factor 2	-0.5502	1				
Factor 3	0.4973	-0.4371	1			
Factor 4	-0.2484	0.2409	-0.3227	1		

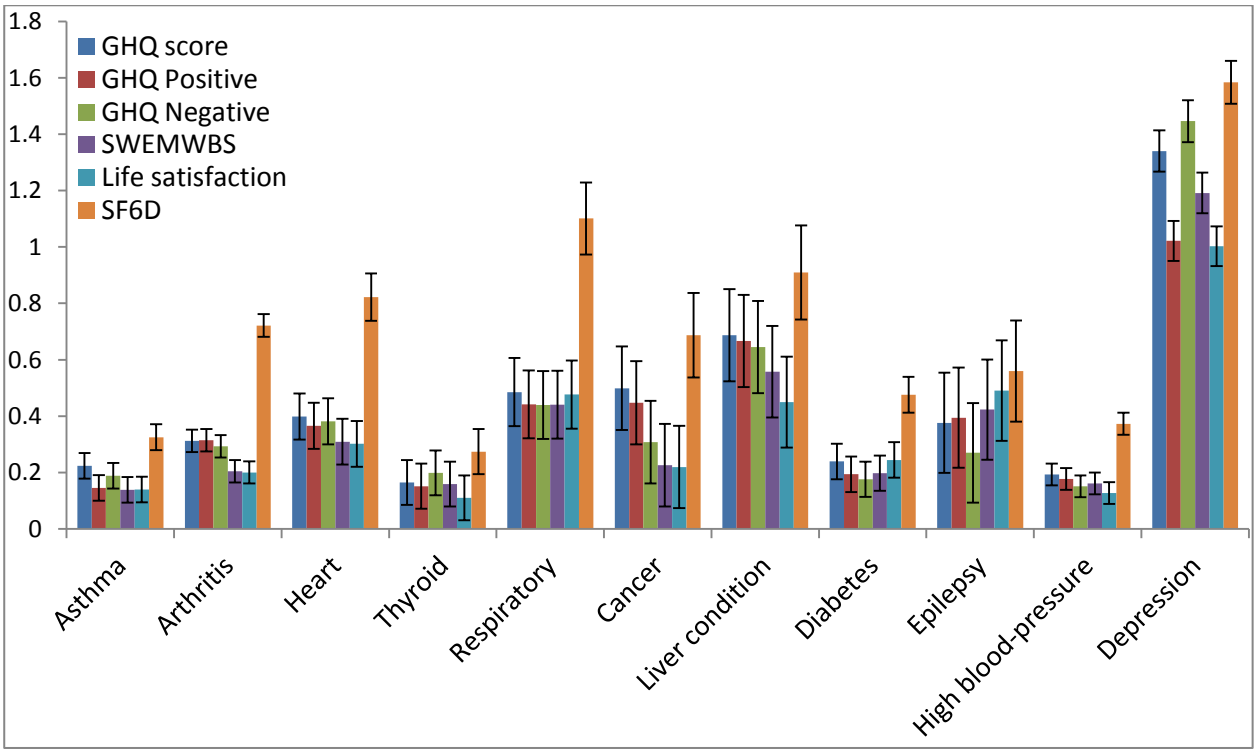
Factor loading |<0.3| are not reported; Extraction Method: Principal Component Factors. Rotation Method: Promax with Kaiser Normalization. Uniqueness >0.6 indicates item may not be contributing to the factor
Cross-loading: factor loading on two or more factors > 0.32

Prefix: ghq – GHQ12 items; swemwbs – Short WEMWBS items;
req – EQ-5D items (recoded)

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). 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, which allow a more accurate control for other important covariates,

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



The SWB Eta^2 effect sizes ranged from 1 to 31% of those for the EQ-5D-5L for different physical health conditions in HIPO (Table 12). There were no mental health conditions in HIPO. There were some differences across the 3 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).

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). As with HIPO, SWB measures ranged from 2 to 8 times as larger as EQ-5D-5L effect sizes.

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 14). The presence of cancer was able to explain more of the variation in happiness compared to EQ-5D-5L (108%). WEMWBS, SWEMWBS 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 14). 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.

In USoc, a similar pattern was observed for GHQ, SWEMWBS 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 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 SWEMWBS 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).

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 size in SWB measures relative to the SF-6D, particularly for the GHQ. 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 confirmed in the fixed effects results where SWB had relatively higher effect sizes for memory and concentration than SF-6D (Table 15d).

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

	satisfaction	worthwhile	happy	anxious (recoded)	ONS-4 total	SWEMWBS	WEMWBS	ICECAP-O	ASCOT	EQ-5D-5L
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
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
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.

Table 15a: 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						
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	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.

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	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.

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

	GHQ	GHQ positive	GHQ negative	SWEMWBS	Full 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 SWEMWBS, 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 measurement error because they are completed in the same way.

Factor analysis indicated that the evaluative and positive affect questions were more likely to be in a single factor. However, the assessment of effect sizes suggested that there were some differences in measures, though this was not consistent across the datasets.

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

There was less overlap between the positive and negative items. The distribution of scores for negative items indicated that there were fewer respondents overall reporting problems in the negative SWB items. Factor analysis suggested that there may be one or two SWB factors related to positive items with a separate factor for negative items. However, this was not conclusive across datasets and may have been influenced by the scaling of measures, for example two positive SWB scales with positive items from the GHQ and SWEMWBS items respectively and one negative GHQ items scale.

From the results one could argue that positive and negative affect (and/or experiences or psychological capacities) are different dimensions of a life. If they are separate, data about one need not necessarily provide information about the other. Furthermore, when they fall into the same dimension they give information about different ends of the spectrum. Knowing about negative affect helps to better understand the SWB of those with low reports of positive affect (or evaluation or psychological capacities). However, it is possible that more detailed psychometric analyses will find that positive and negative affect (or experiences or psychological capacities) fall into the same dimension; even if there are some linguistic or conceptual differences between the absence of happiness and the presence of sadness. Knowledge of one allows the prediction of the other, and this analysis will form a major part of the next report.

Whilst the different factors with the GHQ-12 and the SWEMWBS are clearly identified, it is difficult to judge the extent to which this is driven by differences in instrument layout and response options rather than underlying differences in the concepts. GHQ-12 negative score showed larger effect sizes than the SWEMWBS 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 SWEMWBS, 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.

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.

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. SWEMWBS 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 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 where there were statistically significant changes in SF-6D at follow-up in the USoc data.

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.

- Importantly, none of the datasets contained all the measures of interest. There was overlap between the ONS-4, WEMWWBS, ICECAP-O, other SWB items and health measures or overlap between GHQ-12 and SWEMWBS. Further analysis is required particularly of the SWEMBWS against more conventionally scaled negative SWB questions to address the question of whether or not negative SWB items were required.
- USoc was the only dataset that had follow-up data. 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 which may influence SWB.
- 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). HIPO and SYC65 are self-complete on paper at home, MIC is self-complete online, HSE and USoc SWB and HRQoL are self-compete components that form part of a wider interview which include face-to-face interviews with computer assisted personal interviewing (CAPI). Different collection modes have been found to effect responses to SWB questions. 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] All surveys used rely upon self-report data but there are differences in who is present during the interview. Although these differences are not an issue in comparisons within a dataset, they may limit comparisons across datasets.

- 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.
- Although the datasets were large and mixed, complete case analysis was undertaken which excluded a large number of respondents. 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.
- Responses for HIPO were 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. There is also likely to be a selection effect which cannot be tested for, where those who were doing better responded to the survey. 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 Recommendations for future work

a. The factor analysis conducted here appears to have been influenced by the wording of response options rather than the underlying concept which measurement is aimed at. Further confirmatory factor analysis will be undertaken which allows us to take into account these measurement effects in order to test whether or not they are influencing the results. We will also assess whether or not there are differences across the scale, for example whether the inclusion of negative items improves the measurement precision of factors at the ends of the scale using item response theory methods.
b. We explored the use of a combined measure for the ONS-4 questions, with a simple equal weighting given to the four questions. This was partly because the other measures involved an overall score which would be likely to show greater sensitivity to different health conditions due to a reduction in measurement error arising because of the greater number of questions. However, the anxiety question loaded to a separate factor than the other ONS-4 questions. This suggests that simply combining the scores will be inappropriate. Indeed, the combined score fails to offer any greater sensitivity to health conditions. In future work we will continue to explore the possibility of combining the three remaining positive ONS-4 questions, which may also reduce the random measurement error which tends to be higher in single item measures.

4.4 Implications for policy

a. The results do not provide definite guidance on whether GHQ-12 and SWEMWBS 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. SWEMWBS 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 health conditions and as such would have dramatic 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.

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.

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

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.

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.

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

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.

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, 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.

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.

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.

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

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.

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.

Young 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= 4,783			MIC n=6,808			SYC65 (n=1593)	
	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	

$^{\Phi}$ 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	SWEMWBS	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 $^{\phi}$	SWEMWBS	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	SWEMWBS 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	SWEMWBS	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)

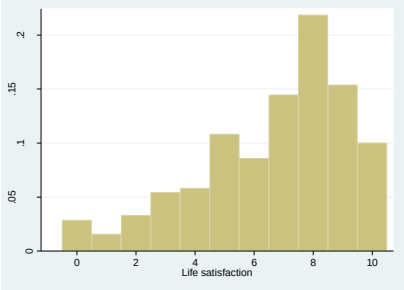
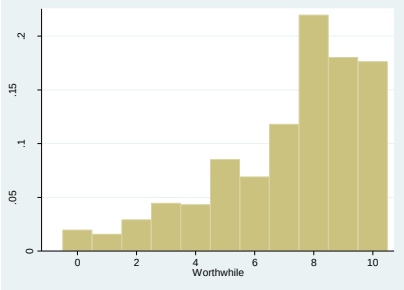
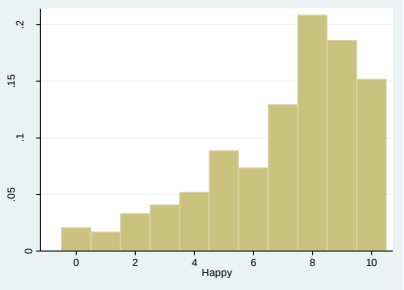
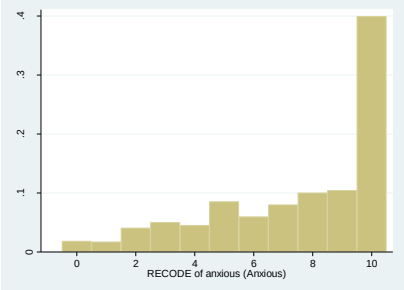
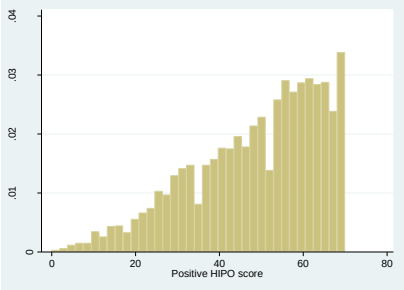
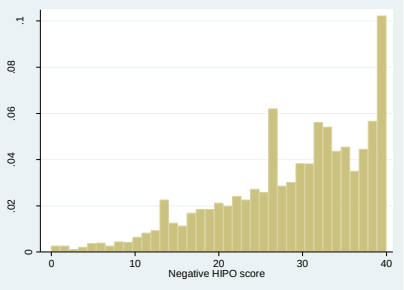
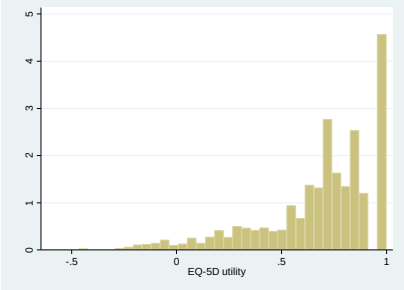
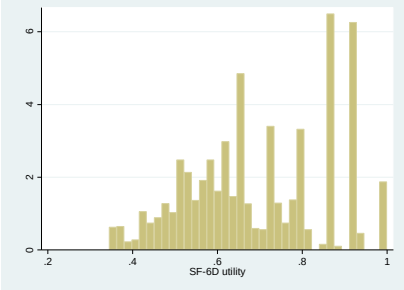
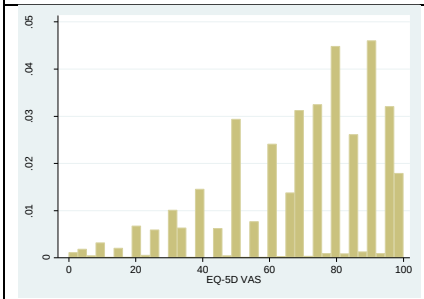
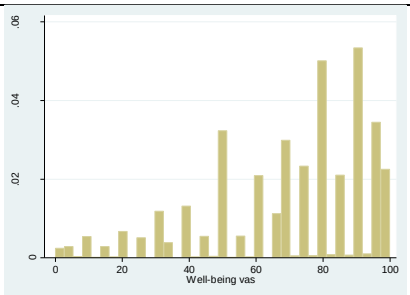
	GHQ $^{\phi}$	GHQ positive $^{\phi}$	GHQ negative $^{\phi}$	SWEMWBS	Full WEMWBS	Happy	EQ-5D-3L
	%	**%**	**%**	**%**	**%**	**%**	**%**
Neoplasms and benign growths	**0.44**	**0.46**	**0.31**	**-0.15**	**-0.19**	**-0.10**	**-0.64**
Endocrine and metabolic	0.01	0.00	0.01	-0.03	-0.04	-0.02	**-0.14**
Mental 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.

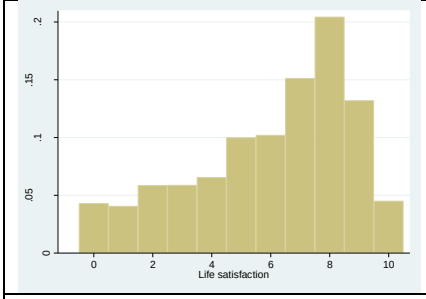
Relative effect sizes calculated where EQ-5D is statistically significant.

Appendix Figure 1 Distribution of SWB and health measures - HIPO

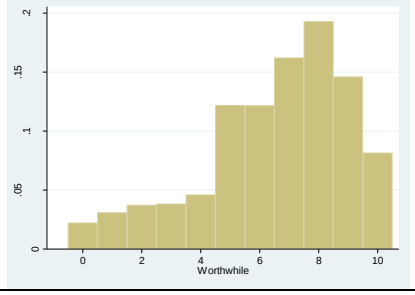
ONS-4 life satisfaction	ONS-4 – worthwhile	ONS-4 happy
		
ONS-4- anxious	**Positive HIPO SWB scores**	**Negative HIPO SWB scores**
		
EQ-5D-5L	**SF-6D (SF-12)**	
		
EQ-5D VAS	**SWB-VAS**	
		

Appendix Figure 2 Distribution of SWB and health measures – MIC

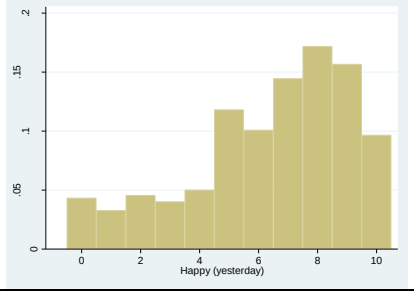
ONS-4 life satisfaction



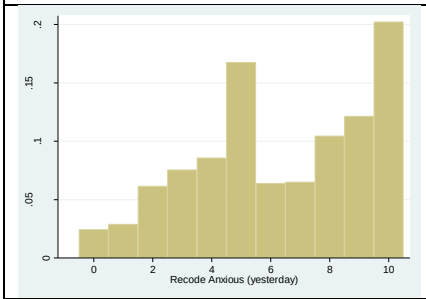
ONS-4 – worthwhile



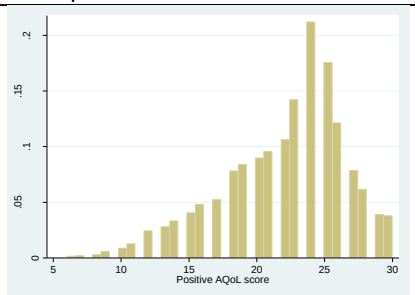
ONS-4 happy



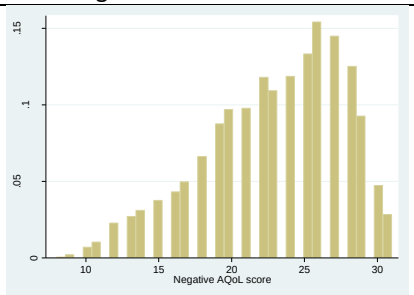
ONS-4- anxious



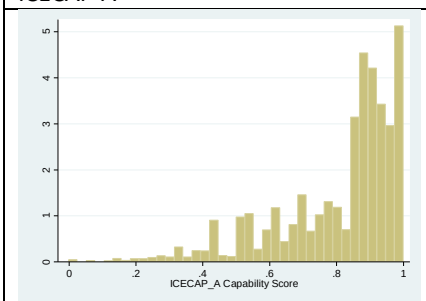
AQoL positive SWB scores



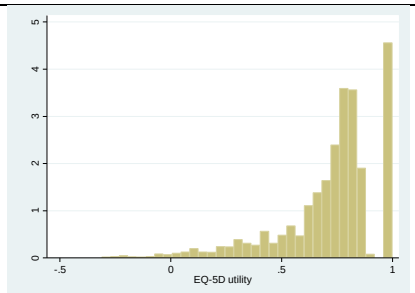
AQoL negative SWB scores



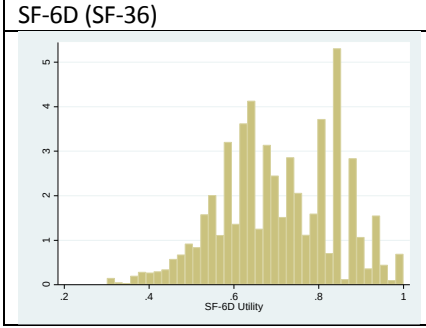
ICECAP-A



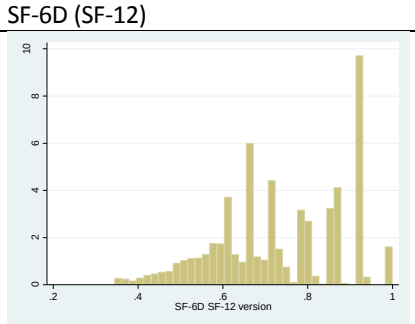
EQ-5D-5L



SF-6D (SF-36)



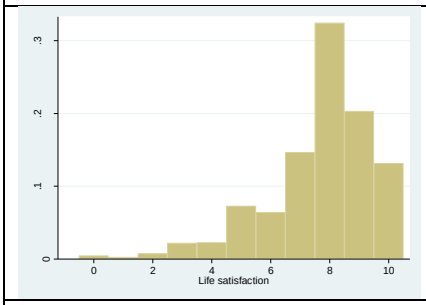
SF-6D (SF-12)



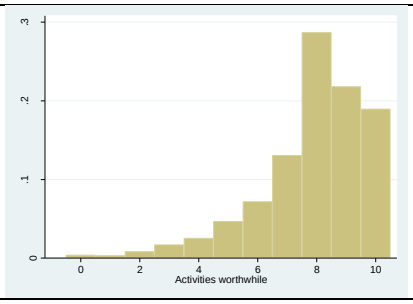
NB: ONS-4 items in MIC had a 'neutral' label at 5

Appendix Figure 3 Distribution of SWB and health measures – SYC65

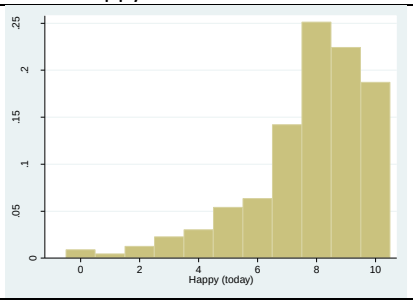
ONS-4 life satisfaction



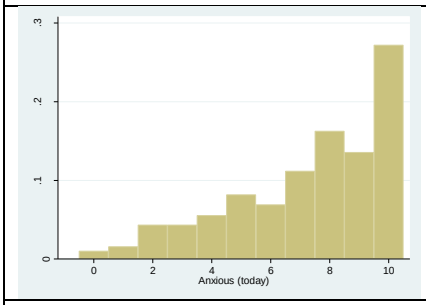
ONS-4 – worthwhile



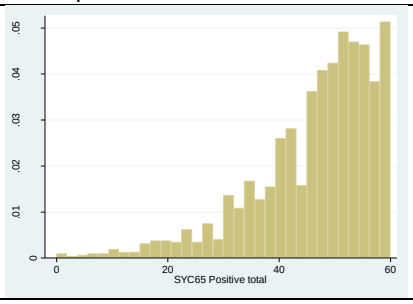
ONS-4 happy



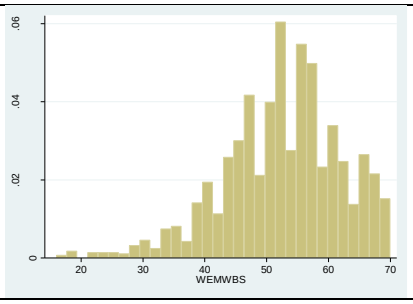
ONS-4- anxious



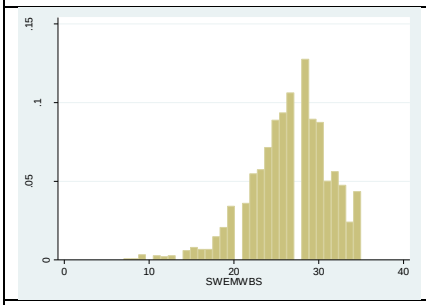
SYC65 positive SWB scores



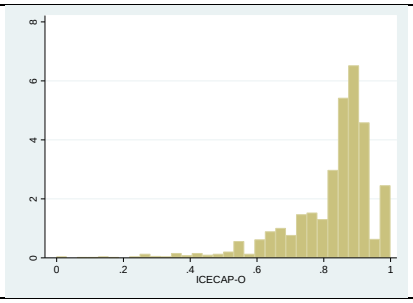
WEMWBS



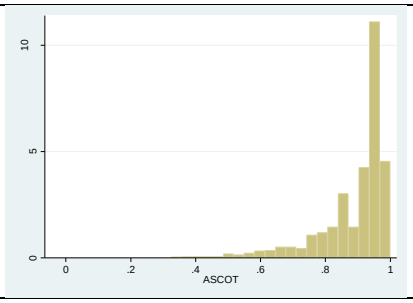
SWEMWBS



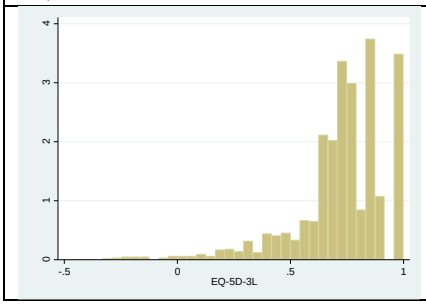
ICECAP-O



ASCOT

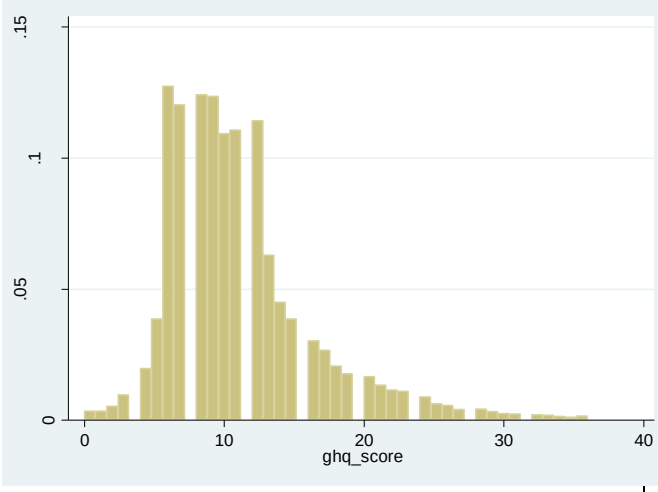


EQ-5D-5L

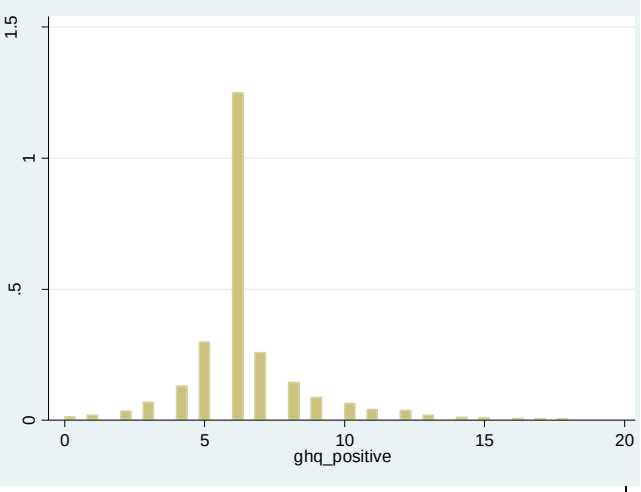


Appendix Figure 4 Distribution of SWB and health measures – Usoc wave 1

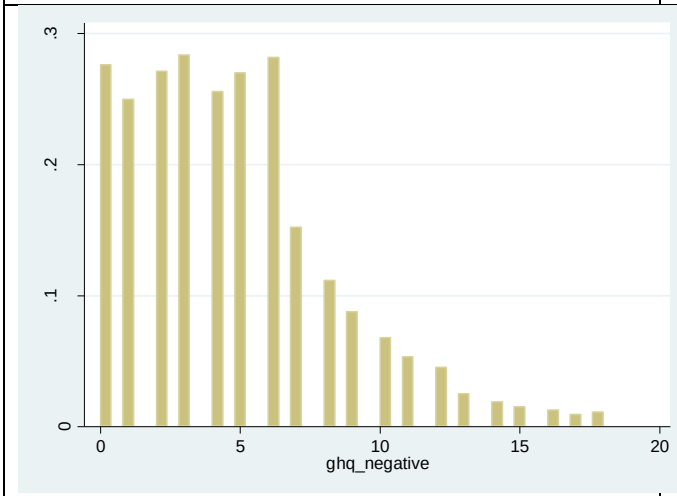
GHQ-12 score



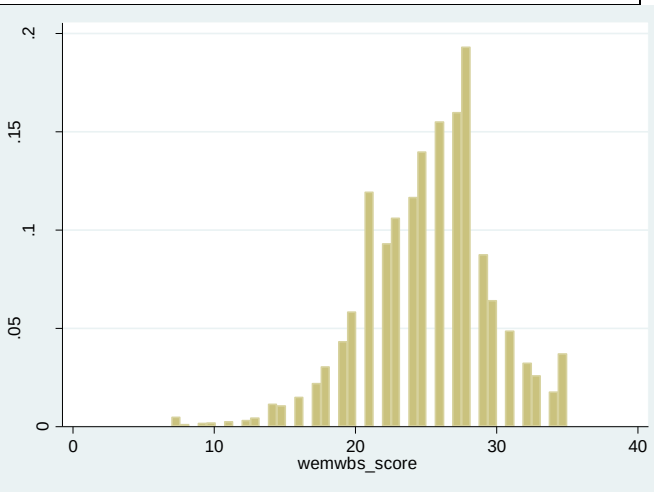
GHQ positive



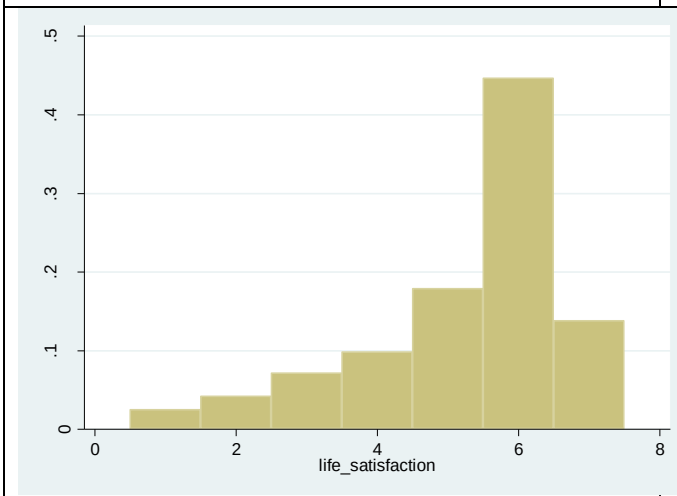
GHQ negative



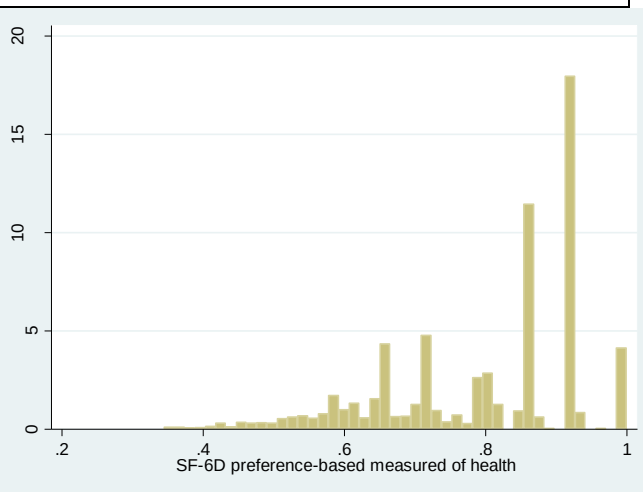
SWEMWBS



Life satisfaction (0 to 7 scale)

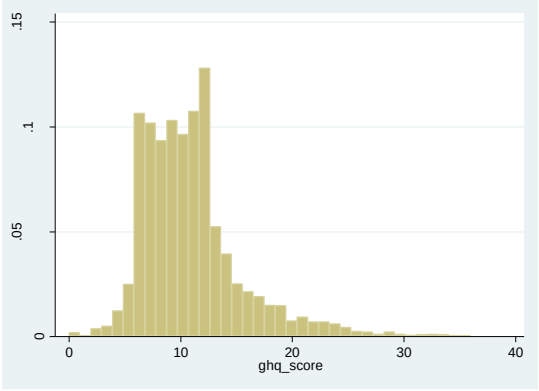


SF-6D (SF-12)

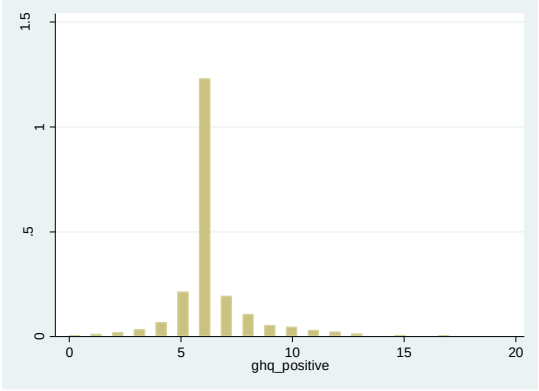


Appendix Figure 5 Distribution of SWB and health measures - HSE

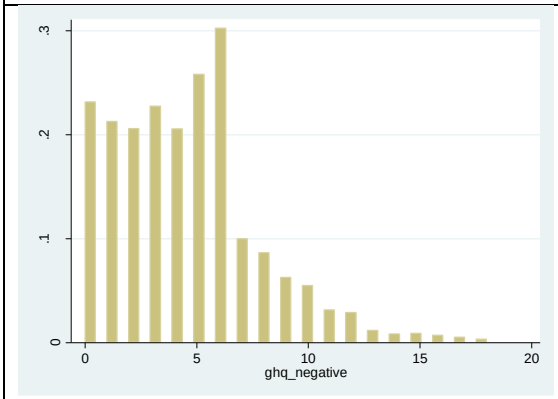
GHQ-12 score



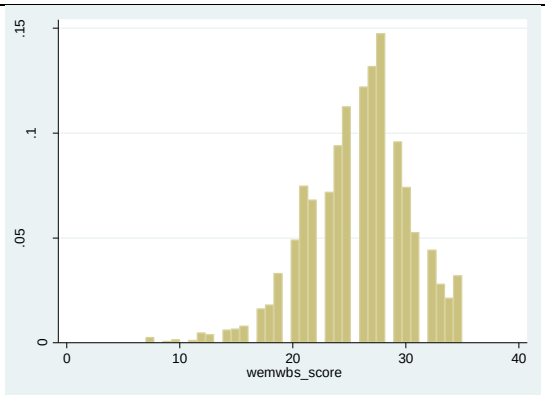
GHQ-positive



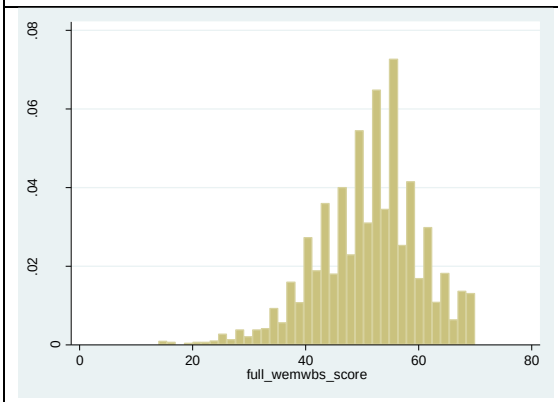
GHQ-negative



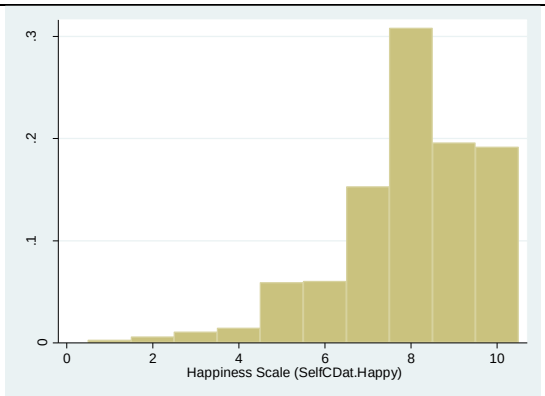
SWEMWBS



WEMWBS



Happiness



EQ-5D-3L