Social Indicators Research
https://doi.org/10.1007/s11205-025-03728-1

ORIGINAL RESEARCH

A Large Scale Population Survey of Health and Wellbeing to Allow Comparisons Between Outcome Measures: the SIPHER-HWMIC Dataset

Nyantara Wickramasekera[1] · **Aki Tsuchiya**[2]

Received: 9 October 2023 / Accepted: 27 September 2024

Abstract
This paper presents a large scale cross-section survey of health and wellbeing that allows econometric mapping between different outcome measures. The Systems science In Public Health and Health Economics Research - Health and Wellbeing Multi-Instrument Comparison (SIPHER-HWMIC) is a cross-sectional individual-level survey conducted in the United Kingdom, of over 12,000 members of the general public aged 18+. It includes a wide range of self-reported questions in health and wellbeing taken from established outcome instruments such as the EQ-5D-5L, Health Utilities Index Mark 3 (HUI3), ICECAP-A, EQ Health and Wellbeing Short (EQ-HWB-S), etc., alongside work-related and housing-related questions, and socioeconomic covariates including age, gender, education, personal income, etc. The data were collected through three separate commercial internet panels in November 2022, and oversampled respondents from Scotland. To demonstrate the potential uses of the dataset, the paper also reports on the results of an econometric mapping analysis between Short Warwick-Edinburgh Mental Wellbeing Scale (SWEMWBS) and the 12-Item Short Form Health Survey (SF-12v2) Mental Component Summary. The SIPHER-HWMIC dataset is available from the UK Data Service (SN9458).

Keywords Patient reported outcome measures · Quality of life · Health · Wellbeing · Mapping · Equivalent income

✉ Aki Tsuchiya
a.tsuchiya@sheffield.ac.uk

[1] Sheffield Centre for Health and Related Research, University of Sheffield, Sheffield S1 4DA, UK

[2] School of Economics and School of Medicine and Population Health, The University of Sheffield, Sheffield S10 2TU, UK

Published online: 09 October 2025

1 Introduction

Traditionally, policymakers have relied on income as a primary indicator of wellbeing. However, this metric is crude as it fails to encompass the multifaceted nature of our lives. A singular focus on income fails to account for a wide range of factors that contribute to our wellbeing, including physical and mental health, social connections, housing quality, and personal security. There is a need to move beyond income by exploring a broader spectrum of health and wellbeing indicators. This requires datasets that collect the same outcome variables. For example, in order to compare mental wellbeing by social class across the four nations of the United Kingdom (UK), data on mental wellbeing needs to be collected using the same measure across the relevant population groups. However, this is perhaps surprisingly difficult. Single UK-wide surveys may not achieve sufficient sample sizes in the smaller nations, while nation-specific surveys may use different outcome measures. This is a gap that we aim to fill. SIPHER-HWMIC (pronounced "*sai-fer hoo-mic"*), stands for Systems science In Public Health and Health Economics Research[1] - Health and Wellbeing Multi-Instrument Comparison, and is a dataset of health and wellbeing indicators across multiple instruments.

SIPHER-HWMIC is a cross-section, individual-level survey of over 12,000 members of the UK public aged 18+. It includes a wide range of self-reported questions in health and wellbeing taken from established outcome instruments such as the EQ-5D-5L, HUI3, ICECAP-A, ONS-4, etc., alongside work-related and housing-related questions, and socio-economic covariates including age, gender, education, household and personal income, etc. The data were collected through three separate commercial internet panels in November 2022, and oversampled respondents from Scotland.

Multi-instrument comparison (MIC) studies have been conducted in health, with adult, patient and paediatric populations (Jones et al., 2021; Mukuria et al., 2016; Richardson et al., 2015). They have been used to assess the psychometric relationships between health related quality of life (HRQoL) instruments (Mukuria et al., 2016; Richardson et al., 2015), and to assess how different instruments that have accompanying preference-based scoring systems (or "value sets" that use common end points of 1 for full health and 0 for being dead) assign different scores to the same outcome. Richardson et al. (2015) conducted a MIC study to assess inconsistencies in values produced by different multi-attribute utility (MAU) instruments. They aimed to understand the underlying dimension of wellbeing captured by the different MAU instruments. The study sought to document these differences across different countries such as Australia, UK, USA, Canada and Norway and across different patient groups using online panels. Jones et al. (2021) conducted a similar study with a specific focus on children's HRQoL, addressing a paucity of information on the psychometric performance of paediatric HRQoL instruments. Recruitment for the study consisted of a hospital-based sample, an online population sample, and an online disease-specific sample. Mukuria et al. (2016) conducted a secondary analysis of five MIC datasets, comparing commonly used measures such as SWEMWBS, GHQ-12, ONS-4, ICECAP, ASCOT,

[1] SIPHER (Systems science In Public health and Health Economics Research) is a consortium of researchers and policy makers that work together to improve wellbeing and reduce inequality by using up-to-date evidence and modelling of complex relationships. It aims to contribute to policy development addressing four key social policy challenges in the UK: inclusive economic growth, housing quality, public mental health, and child poverty (Meier et al., 2019) https://www.gla.ac.uk/research/az/sipher/.

and EQ-5D. Mukuria et al. found that while subjective well-being measures correlate, they differed in capturing different aspects such as positive and negative well-being. Additionally, subjective well-being measures were less sensitive to capturing physical health compared to measures like the EQ-5D.

Such MIC datasets can also be used in mapping between different outcome measures, where an outcome measure of interest is not included in the dataset being analysed. This is a regression-based method often used in health technology assessments, to predict health outcome preference-weights when the patient reported outcome measure (PROM) in a clinical trial is not preference-based (Ara et al., 2017)., Researchers first estimate a mapping function between the PROM and a preference-based instrument using an external dataset (such as a MIC), and then use the regression coefficients to predict the preference-weights for the PROM in the clinical trial (Hernandez et al., 2020; Wailoo et al., 2017; Wolowacz, 2017).

This study expands on existing MIC studies as the focus is not only on comparing HRQoL instruments but also on a preference-based wellbeing measure called equivalent income. Equivalent income translates individual preferences of multi-dimensional wellbeing into a monetary figure that represents the amount of income that, if combined with the best levels in the non-income domains of wellbeing, would be as good as the given situation (Fleurbaey, 2005, 2006, 2009; Fleurbaey & Blanchet, 2013; Fleurbaey & Gaulier, 2009). It is similar to the concept of the Quality-Adjusted Life Year (QALY): the number of years of survival that, if combined with the best levels in the non-survival domains of health, would be as good as the given health prospect.

The aim of this paper is to introduce the SIPHER-HWMIC dataset and to demonstrate its potential for outcomes and indicators research. By compiling a comprehensive dataset incorporating a wide range of health and wellbeing indicators, the objective of this study is to facilitate multi-instrument comparisons. This may lead to the identification of more robust indicators of wellbeing, ultimately supporting the evaluation of policy efficacy and the optimization of resource allocation for improved population health and welfare. The rest of the paper will outline the data collection process; present key descriptive statistics and basic comparisons between instruments for illustration purposes; and demonstrate an example mapping analysis between two mental wellbeing instruments: SWEMWBS and the SF-12v2 MCS.

2 Methods

2.1 The Instruments Used

The SIPHER-HWMIC survey consisted of 84 questions from eight established instruments in health and/or wellbeing, as well as indicators of good work, housing, or neighbourhood, and further socio-demographic questions (Table 1). These survey items were selected after reviewing the relevant literature. Iterative rounds of piloting were conducted with the research team of the wider project to develop the final survey. During piloting some instruments were dropped to reduce respondent fatigue (e.g. AQoL-4D/GHQ-12) and to reduce repetitive items (e.g. SF-6Dv2). The survey questions were all closed ended (e.g. multiple choice) with forced responses to avoid missing data but a "prefer not to say" option was provided where possible (e.g. EQ-5D-5L and SF-12v2 do not allow this).

Table 1 Survey item descriptions

Instruments	Domain covered	No of items	Recall period	Outcomes	Scoring
SIPHER-7	Wellbeing	7	Past 4 weeks (physical health and emotional items) rest none	effect of physical health, effect of emotional problems, loneliness, disposable household income, employment, quality of home, safety of neighbourhood	Preference-based index
EQ-5D-5L	Health	6	Today	mobility, self-care, usual activity, pain, anxiety/depression, visual analogue scale	Preference-based index and VAS score
SF-12v2	Health	12	Past 4 weeks	overall health, mobility, physical health, mental health, pain, calm, energy, downhearted	Standardised summary scoring for physical and mental health
HUI3	Health	8	Usual	vision, hearing, speech, ambulation, dexterity, emotion, cognition, pain	Preference-based index
ICECAP-A	Capabilities	5	At the moment	stability, attachment, autonomy, achievement, enjoyment	Preference-based index
EQ-HWB-S	Health and wellbeing	9	Last 7 days	mobility, activity, exhausted, lonely, concentrate, anxious, depressed, control, pain	Preference-based index
ONS-4	Wellbeing	4	Overall/yesterday	life satisfaction, worthwhile, happiness, anxiety	NA
SWEMWBS	Wellbeing	7	Last two weeks	optimistic, useful, relaxed, dealing with problems, thinking clearly, feeling close to people	Total score
Good work indicators	Work	7	-	meaningful work, guaranteed work hours, shift work notice period, minimum living wage, job satisfaction	NA
Housing/neighbourhood quality indicators	Housing	11	-	state of repair, warm in winter, damp-free home, vandalism, break-ins, attacks in the streets	NA
Sociodemographic	-	8	-	Age, gender, education, personal income, household size, ethnicity, dependents, caring responsibilities	NA

HUI3: Health Utilities Index Mark 3; EQ-HWB-S: EQ Health and Wellbeing Short version; SWEMWBS: The Warwick-Edinburgh Mental Wellbeing Scale short version; ICECAP-A: ICEpop CAPability measure for Adults

SIPHER-7 is a wellbeing outcome measure with seven items (Tsuchiya & Wu, 2021). Neighbourhood safety, Housing, and Loneliness have three ordered levels each; Effects of physical health and Effects of mental health have five ordered levels; and Employment has nine (unordered) categories. Income is defined as household disposable income after housing costs and can be treated as continuous. However, since measurement error can be large, especially in online surveys, we have offered six ranges based on descriptive statistics taken from the UK Household Longitudinal Study (UKHLS; University of Essex, 2020) for respondents to select from. The items of SIPHER-7 are linked to questions used in the UKHLS so that it is possible to determine a SIPHER-7 profile for UKHLS observations (Tsuchiya et al., 2022). The scoring system is preference-based, consisting of equivalent income that penalises household disposable income (equivalised for household size) to adjust for non-income items that are not at the best levels (Wickramasekera et al., forthcoming).

EQ-5D-5L is a health outcome measure with five items, each with five ordered levels (Herdman et al., 2011). The scoring system is preference-based, consisting of "utility" weights that are anchored on 1 for full health and 0 for being dead, and we use the algorithm estimated for England (Devlin et al., 2018). EQ-VAS is part of the EQ-5D instrument (Brooks and the EuroQol Group, 1996) and asks the respondent to indicate how good or bad "your health today" is on a vertical scale between 0 for "worst imaginable health" and 100 for "best imaginable health".

SF-12v2 is a health outcome measure with 12 items, each with five ordered levels, except two items with three ordered levels (Ware et al., 2002). The 12 items are clustered into the eight scales including physical functioning, role physical, general health, bodily pain, vitality, social functioning, role emotional, and mental health. SF-12v2 produces mental and physical component summary scores, which are not preference-based and gives each relevant item equal weight. These can be standardised using the 2009 U.S. general population t-scores where the standardised scores have a mean of 50 and a standard deviation of 10, with higher scores representing better physical and mental health (Ware et al., 2005). Using a subset of the 12 items, a preference-based measure SF-6Dv1 can be used to derive utilities anchored on 1 for full health and 0 for being dead (Brazier and Roberts, 2004).

HUI3 is a health outcomes measure with eight items, each with between five and six levels. The eight items cover functional (dis)ability in the following areas: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain. The scoring system is preference-based, anchored on 1 for full health and 0 for being dead using a Canadian value set (Feeny et al., 2002).

ICECAP-A is a capability measure that contains five attributes each with four levels (Flynn et al., 2015). The attributes are stability, attachment, autonomy, achievement, and enjoyment. The scoring system is preference-based using a UK value set. A utility score of 1 is equal to full capability and 0 for no capability (Flynn et al., 2015).

EQ-HWB-S is a health and wellbeing measure that contains nine items, each with five ordered levels. The eight items are mobility, activity, exhausted, lonely, concentrate, anxious, depressed, control and pain. The scoring system is preference-based with a score of 1 is equal to perfect health and wellbeing, and 0 is equal to being dead, using a UK value set (Mukuria et al., 2023).

ONS-4 is a wellbeing measure that contains four questions, each on a scale from 0 to 10. The measure asks people to evaluate how satisfied they are with their life, whether they feel

worthwhile, and their emotional state – happiness and anxiety (ONS, 2018). There are no total scores, and sample means are reported for each question separately. ONS also provides cut offs scores, 9–10 grouped as Very high, 7–8 as High, 5–6 as Medium and 0–4 as Low for the life satisfaction, worthwhile and happiness items, and 6–10 as high, 4–5 as medium, 2–3 as low, and as 0–1 very low for the anxiety item (ONS, 2018).

SWEMBS is a wellbeing measure that contains seven items, each with five ordered levels (University of Warwick, 2011). The seven items are optimistic, useful, relaxed, dealing with problems, thinking clearly, and feeling close to people. The seven items are scored on a 5-point Likert scale, from none of the time receiving a score of 1 to all of the time receiving a score of 5. The total score is between 7 and 35 where higher scores indicate higher levels of wellbeing.

In addition to the eight outcome measures detailed above, further questions were included in the survey about good work indicators, quality of housing, safety of neighbourhood, and sociodemographic questions (See Table 1 and the survey in the Supplementary material). These questions were all taken from existing surveys such as the UKHLS, English Housing Survey, Labour Force Survey, and the Living Wage Foundation Survey.

2.2 Data Collection

2.2.1 The Survey

The survey consisted of two parts. Before taking part respondents reviewed an information sheet and provided informed consent. This was followed by the main section with the health and wellbeing instruments. The order in which the instruments were displayed was randomised to minimise any order bias, but ordering of items within a given instrument was fixed to maintain the integrity of the outcome measure. Minimum completion time cut-offs were not introduced, and respondents were allowed to stop and come back to the survey up to 72 hours after clicking on the link.

2.2.2 The Sample and Recruitment

Residents living in the UK above the age of 18 were eligible to take part in the study, with no upper age limit. Quotas were introduced to recruit a representative sample of the UK in terms of age and sex. Additional sampling targets were applied to recruit a 50:50 split between Scotland and the rest of the UK to allow subgroup analyses for the Scottish sample. The target sample size was 10,000 respondents and this number was selected based on a similar multi-instrument comparison study recruiting over 9000 + respondents (Richardson et al., 2015).

The survey was administered by three commercial internet panels – Prolific (UK provider), MRFGR (UK provider) and STV (Scottish provider). Three panels were used given the large sample size requirement of 10,000 respondents, and to allow comparisons across panels. For the MRFGR and Prolific panels, a Qualtrics survey was set up by the research team, and a link to it was sent by the panels to eligible members, inviting them to take part in the survey. For the STV panel, STV coded the same survey on their platform (Forsta) and invited eligible panel members.

The survey was approved by the University of Sheffield's School of Health and Related Research Ethics Committee (ref: 049491).

2.3 Data Analysis

2.3.1 Descriptive Statistics

Descriptive statistics were calculated for all survey items and the derived summary scores using frequencies and percentages for categorical variables, and means, standard deviations, medians, interquartile ranges for continuous variables. Group differences were explored at the nation level (Scotland vs rest of the UK) and by panel providers (MRFGR vs Prolific vs STV) using Chi^2 tests.

To adjust for sampling differences in age and sex, a sampling weight variable was created, utilising population statistics taken from 2021 Mid-Year Population Estimates for the UK (ONS, 2022). Sample weights were included for the following sociodemographic variables: age, gender, education, personal income, household size, ethnicity, dependents, and caring responsibilities.

Data quality was assessed in three ways. First, the time it took respondents to complete the survey was recorded as percentiles and medians (IQR). Second, we explored the number of respondents selecting the "prefer not to say" option at any point during the survey. Third, proportion of those who are inconsistent when the same question is flipped was assessed making use of two questions that were repeated in the survey. The response options for the physical and emotional functioning questions in SF-12v2 go from 'All of the time' to 'None of the time', whereas the response options for the same two questions in SIPHER-7 start from 'None of the time' with 'All of the time' last.

All these data quality variables were analysed by grouping the respondents into panel provider subgroups using Chi^2 tests.

2.3.2 Derived Variables

In addition to the variables collected in the online survey, existing scoring systems were applied where relevant (see Table 1) to derive summary scores.

Preference-based health utility scores anchored at 1 for full health and 0 for being dead were calculated for the three health instruments, using published algorithms: EQ-5D using the English value set (Devlin et al., 2018); SF-6Dv1 using the UK value set (Brazier & Roberts, 2004); and HUI3 using the Canadian value set (Feeny et al. 2002). In addition, for EQ-HWB-S, preference-based utility scores anchored at 1 for full health and wellbeing and 0 for being dead were calculated using Mukuria et al (2023).

Preference-based scores with alternative anchors were calculated for three further instruments. The EQ-VAS score gives a preference-based representation of the respondent's health, but the lower anchor is set at the worst imaginable health state (which may be better or worse than being dead). For the capability instrument, ICECAP-A, the anchors are 1, equal to full capability, and 0, equal to no capability (Flynn et al., 2015).

The scoring system of SIPHER-7 is based on equivalent income. This was calculated using an algorithm that was estimated from a Discrete Choice Experiment survey of the UK public (Wickramasekera et al., forthcoming), and collapses the seven items of SIPHER-7 by

using the six non-income items to penalise the income item (see Supplementary material). To calculate equivalent income a point estimate of household income is needed. However, in the SIPHER-HWMIC survey participants self-reported their household income band, which were: up to £690 per month; up to £1040 per month; up to £1380 per month; up to £1730 per month; up to £2080 per month; and higher than £2080 per month. For the main analysis, we use the midpoint of the range, with the midpoint for the last category (higher than £2080 per month) estimated from the UKHLS survey. The supplementary material reports an alternative method using an interval regression to predict an estimated household income based on sex, age, household size, education, personal income, dependents and caring responsibilities as explanatory variables.

Non-preference-based summary scores were calculated for two instruments. The SF-12v2 physical and mental component summary scores were calculated using Quality Metrics Pro CoRE program (Ware et al., 2002). Simple additive scoring was applied across the seven items of SWEMWBS (University of Warwick, 2011). And finally, the ONS-4 items were presented separately, since they have no summary scoring system to aggregate across them (ONS, 2018).

2.3.3 Comparison of Health Related Quality of Life and Wellbeing Instruments

To provide an overview of the health and wellbeing instruments, associations between the following were explored: SIPHER-7, EQ-HWB-S, EQ-5D-5L, EQ-5D-VAS, HUI3, SF6Dv1, ICECAP-A, SWEMWBS, the Mental Component Score of SF-12v2, the four items of ONS-4, and the Physical Component Score of SF-12v2. This was done by calculating the Pearson correlation coefficients across the summary scores for each pair of instruments. The scale of the correlation coefficients range from −1 to + 1, where 0 indicates no association, < (±)0.1 a negligible correlation, < (±)0.4 a weak correlation, < (±)0.7 a moderate correlation, < (±)0.9 a strong correlation and > (±)0.9 a very strong correlation (Schober et al., 2018).

2.3.4 Mapping Between SF-12v2 MCS and SWEMWBS

As a final example, regression analyses were used to explore the relationship between the SF-12v2 mental component summary score (MCS) and the seven items of SWEMWBS. Mapping is an econometrics-based technique that is widely used in health economics, and in health technology assessment in particular, where, for example, a typical mapping study might map from a condition-specific health outcome measure that is not preference-based to a generic preference-based health outcome measure, in order to allow preference-based analyses of a trial dataset that does not include a preference-based measure (Petrou et al., 2015; Wailoo et al., 2017). Here, we map between two mental health outcome measures, neither of which is preference-based. The dependent variable, SF-12v2 MCS, is one of the two summary scores of SF-12v2 that adds across the items on vitality, social functioning, role-emotional and mental health. The explanatory variables of the mapping model are all seven SWEMWBS items.

Four models were estimated with different sets of explanatory variables: 1) all seven items of SWEMWBS each with all five levels treated as categorical (viz. four dummy variables per item); 2) all seven items of SWMWBS but collapsing all non-ordered/monotonic

coefficient levels within a given item; 3) model 2 plus a dummy variable for selecting the best level ('all of the time') three or more times across the SWEMWBS items to account for any ceiling effects; and 4) model 3 plus age and sex dummies. All models assume that item dimensions are additive and use ordinary least squares (OLS). The results are presented along with goodness of fit statistics as well as scatter plots of model predictions compared with the dependent variable.

3 Results

3.1 The Sample and Response Rates

This online survey was carried out between November and December 2022. The total number of respondents completing the survey was 12,401 (Table 2). The completion rate (number completed/number consented) was 83.4% and the response rate (number completed/number entering the survey) was 75.7%. We exceeded the original sample size estimate of 10,000 respondents because quotas were not strictly implemented in the STV panel, and consequently, the survey was kept open to recruit under sampled respondents (males and younger respondents). Those who partially completed the survey (and therefore not included in the final sample) compared to those who fully completed the survey were no different in terms of gender. However, the dropouts were significantly more likely to be in the age groups 18–25 and 26–45 (Table 2).

3.2 Participant Characteristics

The England, Wales and Northern Ireland sample is broadly representative of the general public in terms of gender and age (Table 3). However, the Scottish sample is represented

Table 2 Response and completion rate

Panel	Number fully completed the survey	Number consented to take part in survey	Number entered the survey and clicked on information sheet or consent form
STV	7366	8464	9826
Prolific	1501	1513	1515
MRFGR	3534	4895	5034
Total	12,401	14,872	16,375
Comparison of complete responses vs partial responses^			
Gender:	**Complete %**	**Partial %**	**Chi^2 (p-value)**
Male	49.1	48.4	2.70 (0.259)
Female	50.6	50.5	
Other	0.3	1.1	
Age by category:			
18 to 25	13.4	20.9	8.80 (0.032)
26 to 45	33.1	28.2	
46 to 65	35.1	32.2	
66 +	18.4	18.6	

Note: ^This comparison is made excluding STV data as partial data were only available for MRFGR and Prolific panels

by slightly more females compared to males and a higher number of people in the older age categories compared to the Scottish census (Table 3). Across the whole sample, the majority of the respondents lived in two-person households, were educated to university degree level, were from a white ethnic background, and had no childcare or carer responsibilities. Some participant characteristics were significantly different depending on the panel the participants belonged to and by location (Table 3). Respondents in the Prolific panel were more likely to have a university degree and less likely to be older (66 or above). STV panel members were all from Scotland, more likely to be female, older (45 years and above), or retired.

3.3 SIPHER- 7

The majority of the sample reported no physical health and emotional problems, was in full time employment, lived in good housing and safe neighbourhoods (Table 4). Further details on good work indicators, housing quality and neighbourhood safety are presented in supplementary materials (Table 16 and Table 17). Approximately half the sample reported feeling lonely and left out from others. While all the disposable household income levels were represented by the respondents completing the survey, the majority were in the lower bands, with a quarter (26%) of the respondents reporting disposable household income levels of less than £690 per month. In addition, 16% of the respondents selected the prefer not to say option for disposable household income. Across the questions, there were significant differences by sex and panel provider. For example, the proportion of people reporting emotional problems "all of the time" was significantly higher in the 18–25 age group compared to the older age categories.

3.4 Data Quality

First, the median time to complete the survey was 13 minutes (Table 5). Second, the majority (75.7%) of respondents answered the survey without selecting the prefer not to say option. However, a large proportion (13%−16%) opted for the prefer not to say option for either disposable household income or personal income variables (Tables 3-4). Third, the proportion of those who are inconsistent when the same question is flipped was 48.5% for the physical health question and 47.2% for the emotional health question (Table 5, and Supplementary Table 19). There were also significant differences based on panel provider with Prolific panel members significantly more likely to be consistent and MRFGR panel members less likely to be consistent with both items. Furthermore, MRFGR panel members were more likely to complete the survey in under 6 minutes.

3.5 Derived Variables

For the whole sample, the median equivalent income derived from SIPHER-7 was £313.38 per month. Equivalent income for the four most frequently observed SIPHER-7 profiles is presented in Table 6. The most common profiles of SIPHER-7 are having the best state in the non-income domains with different categories of employment. For example, at the best non-income domains, fully employed people have a higher median equivalent income of £1099.99, compared to part-time employees who have £725.17 or retired respondents who

have £552.24. When respondents reported feeling lonely some of the time but were in the best state for the rest of the domains, the median equivalent income was £794.21 per month.

The mean utility for EQ-HWB-S, EQ-5D-5L, HUI3, SF-6Dv1, ICECAP-A are broadly similar in the range of 0.71 to 0.82, indicating that the sample was generally in good health (see Table 7). SF-12v2 PCS has an approximate mean of 50 and SD of 10 indicating that our sample is similar to the 2009 US benchmark. The MCS is 45 with a SD of 11 indicating that our sample has worse mental health compared to the 2009 US benchmark. The SWEMWBS mean score of 24.15 is similar to the UK population norms mean of 23.61 in 2011 (University of Warwick, 2011).

3.6 Comparison of Utility Tariffs and Total Scores of Instruments

Pearson's correlation coefficients across the 14 summary measures of the five instruments are shown in Fig. 1 and summary statistics are reported in the supplementary materials (Table 9—Table 15). Two groups of summary measures with strong correlations (>0.7) can be observed, one amongst the measures capturing overall health (EQ-HWB-S, EQ-5D-5L, EQ-5D-VAS, HUI3 and SF-6Dv1), and the other amongst the measures capturing mental wellbeing (ICECAP-A, SWEMWBS, SF-12v2MCS, and three of the ONS4 indicators). Notable are the inclusion of EQ-HWB-S in the former group despite its stated focus on health and wellbeing; and the exclusion of ONS4-anxiety from the latter group, possibly caused by the flipped answer categories. Equivalent income is weakly correlated with all the instruments, indicating that the health and wellbeing constructs measured by other instruments have limited overlap with equivalent income. The SF-12v2 MCS and PCS are not correlated with each other.

3.7 Mapping Between SF-12v2MCS and SWEMWBS

Figure 2 is a scatterplot between SF-12v2MCS and the SWEMWBS summary score. The plot illustrates that, while there is an overall positive correlation between the two instruments, the range of SF-12v2MCS, given a SWEMWBS score, is wide. For example, the modal SWEMWBS score is 28, with 1087 respondents, and their SF-12v2MCS range from 19 to 67. The red line is the linear regression prediction of SWEMBS score based on SF-12v2 MCS.

Table 8 shows the results of the mapping regressions between SF-12v2 MCS and SWEMWBS items at the individual level. Model 1 shows the regression results that included all the SWEMWBS item levels as categorical dummies. The items that consistently performed well (e.g. significant and positive coefficient) were feeling relaxed and feeling close to other people. The two domains – making my own mind up and thinking clearly—were disordered and had unexpected negative coefficients. Further analyses by panel show that this disordering is mostly attributable to MRFGR panel respondents (results available upon request). Model 2 shows regression results where the disordered item levels were collapsed (this results in a disordering of dealing with problems domain, but we have not collapsed this further). Model 3 further included an explanatory variable of being in the best level for the SWEMWBS for three or more items to control for ceiling effects (this results in a disordering in feeling useful, but we have not collapsed this further). Model 4 is the preferred model which includes the collapsed SWEMWBS items, the ceiling dummy, and age bands

Table 3 Participant demographic characteristics by UK location and panel provider

N=12,401	Subgroups by: location^		Subgroups by: panel provider^			Overall sample (un-weighted)	Overall sample (weighted)
	Scotland	England, Wales and Northern Ireland	MRFGR	Prolific	STV		
	%	%	%	%	%	%	%
Gender:							
Male	41.2	48.9	49.3	48.4	40.7	44.1	48.4
Female	58.4	50.7	50.5	51.0	58.9	55.6	51.6
Other	0.4	0.3	0.2	0.7	0.4	0.4	
Where do you live:							
England	0.0	91.3	83.0	85.9	0.0	34.1	37.1
Scotland	100.0	0.0	7.8	9.0	100.0	62.7	59.5
Wales	0.0	5.8	6.2	3.3	0.0	2.2	2.2
Northern Ireland	0.0	2.9	3.0	1.7	0.0	1.1	1.2
Age by category:							
18 to 25	4.8	13.6	13.9	12.2	4.4	8.1	12.0
26 to 45	23.7	33.0	30.9	38.4	23.1	27.2	33.3
46 to 65	46.2	35.0	33.7	38.4	46.7	42.0	32.5
66+	25.3	18.4	21.5	11.1	25.7	22.7	22.2
How many people are there in your household:							
I am living alone	20.3	18.1	19.4	16.4	20.1	19.4	19.0
2 people	49.6	38.5	38.3	39.2	50.2	45.5	43.1
3 people	16.0	19.1	18.8	19.8	15.8	17.2	17.8
4 people	10.5	15.4	14.5	16.8	10.4	12.4	13.6
5 or more people	3.2	8.3	8.3	7.5	3.1	5.1	6.1
I prefer not to answer	0.3	0.6	0.7	0.3	0.3	0.4	0.5
Highest educational qualification:							
No formal qualification	6.7	3.1	3.9	1.3	6.8	5.3	4.9
GCSE or equivalent	16.7	20.1	21.1	16.1	16.8	18.0	17.4
A-level or equivalent	19.2	23.9	24.4	22.9	18.9	20.9	21.6

Table 3 (continued)

N=12,401	Subgroups by: location^		Subgroups by: panel provider^			Overall sample (un-weighted)	Overall sample (weighted)
	Scotland	England, Wales and Northern Ireland	MRFGR	Prolific	STV		
	%	%	%	%	%	%	%
University degree (undergraduate/postgraduate)	35.9	45.4	41.5	55.5	35.2	39.4	40.7
Other qualification	19.8	6.9	8.3	4.1	20.4	15.0	14.0
I prefer not to answer	1.8	0.6	0.8	0.2	1.8	1.3	1.3
Total personal income (before deductions including benefits):							
NO INCOME AT ALL	5.2	5.4	5.5	5.7	5.2	5.3	5.3
up to 10,800	15.5	19.0	18.4	20.3	15.4	16.8	16.7
up to 14,400	11.7	10.2	10.6	9.7	11.7	11.1	10.8
up to 19,200	10.2	11.4	11.7	10.3	10.2	10.6	10.8
up to 26,400	14.7	16.3	15.4	18.1	14.6	15.3	15.4
up to 37,200	12.2	14.2	13.9	15.9	11.9	13.0	13.4
up to 52,800	8.4	9.0	8.6	9.3	8.5	8.6	9.0
up to 74,000	3.1	3.2	3.4	2.6	3.1	3.1	3.3
74,000 or more	1.9	2.4	2.7	1.5	1.9	2.1	2.2
I prefer not to answer	17.1	8.8	9.9	6.7	17.5	14.0	13.3
Ethnicity:							
White	97.8	86.1	86.3	87.4	98.0	93.4	92.1
Black, Black British, Caribbean or African	0.2	2.9	2.7	2.9	0.2	1.2	1.5
Asian or Asian British	0.5	6.7	6.1	7.2	0.3	2.8	3.5
Mixed or multiple ethnic groups	0.5	2.3	2.7	1.3	0.4	1.2	1.5
Other ethnic group	0.2	1.0	1.0	0.7	0.2	0.5	0.6
I prefer not to answer	0.9	1.0	1.1	0.6	0.9	0.9	0.9
Do you have responsibility for any dependent children aged under 18:							
Yes	17.7	25.5	25.5	24.5	17.5	20.6	22.3
No	82.0	73.9	73.7	75.3	82.2	79.0	77.2
I prefer not to answer	0.3	0.6	0.8	0.2	0.3	0.4	0.5

Table 3 (continued)

N= 12,401	**Subgroups by: location^**		**Subgroups by: panel provider^**			**Overall sample (un-weighted)**	**Overall sample (weighted)**
	Scotland	**England, Wales and Northern Ireland**	**MRFGR**	**Prolific**	**STV**		
	%	**%**	**%**	**%**	**%**	**%**	**%**
Do you look after anyone with long-term illnesses/problems related to old age:							
No	77.8	80.2	78.7	84.3	77.5	78.7	79.6
Yes, 9 h a week or less	9.2	7.4	7.4	7.3	9.3	8.5	8.0
Yes, 10 to 19 h a week	2.9	2.9	3.0	2.6	2.9	2.9	2.8
Yes, 20 to 34 h a week	1.7	2.2	2.7	1.0	1.7	1.9	1.9
Yes, 35 to 49 h a week	1.9	2.1	2.5	1.1	1.9	2.0	1.9
Yes, 50 or more hours a week	5.1	3.0	3.2	2.5	5.2	4.3	3.9
I prefer not to answer	1.5	2.2	2.4	1.3	1.5	1.7	1.9

Note: ^ difference between the subgroups are significant based on chi^2 test with $p<0.001$ for all demographics reported in Table 1. Column % add up to 100 for each subcategory (e.g. Gender)

Table 4 SIPHER-7 by sex, age, and panel

	Subgroups by: sex			Subgroups by: age					Subgroups by: panel provider				Overall sample
	Male	Female	Chi^2	18 to 25	26 to 45	46 to 65	66+	Chi^2	MRFGR	Prolific	STV	Chi^2	
	%	%	Signif	%	%	%	%	Signif	%	%	%	Signif	%
Accomplished less because of your physical health (in the past 4 weeks):													
None of the time	42.5	39.2	16.26	37.7	45.1	40.9	35.9	111.34	38.7	43.3	41	59.85	40.6
A little of the time	23.6	24	**	25.5	23.3	22.7	26	***	23.7	27.1	23.2	***	23.8
Some of the time	18.2	19.1		20.1	17	18.6	20.5		19	17.7	18.8		18.7
Most of the time	10.1	11.5		9.7	8.9	11.4	12.7		11.2	9	11.1		10.9
All of the time	4.9	5.4		5	4.9	5.8	4.4		6.3	2.8	5.1		5.2
I prefer not to answer	0.7	0.7		2	0.8	0.6	0.5		1	0.1	0.7		0.7
Accomplished less because of your emotional problems (in the past 4 weeks):													
None of the time	50.5	40.4	135.07	16.9	31.9	48.7	62.9	1068.59	43	42.6	46.1	62.8	44.8
A little of the time	21	24.6	***	26.1	27.9	22.2	17.5	***	21.6	27.2	22.8	***	23
Some of the time	15.9	19.8		27.3	22.9	16.5	11.9		19.4	17.9	17.4		18
Most of the time	8.3	10.8		20.4	11.7	8.8	5.4		10.1	9.1	9.7		9.8
All of the time	3.4	3.9		7.9	4.8	3.4	1.3		4.7	2.9	3.3		3.7
I prefer not to answer	0.8	0.6		1.4	0.8	0.4	1		1.2	0.2	0.6		0.7
You feel lonely and left out from others:													
Hardly ever	53.7	44.1	120.65	20.5	35.8	52	66.1	973	46.4	49.4	48.9	23.35	48.3
Some of the time	31.2	37.7	***	46.9	42.3	32.8	25.3	***	35.2	33.2	34.9	***	34.8
Often	14.2	17.5		30.3	21	14.6	8		17.1	16.9	15.5		16.1
I prefer not to answer	1	0.7		2.3	0.8	0.7	0.5		1.3	0.5	0.7		0.8
Disposable household income after housing costs:													
Up to 690 per month	23.7	28.8	187.02	27	27.5	28.4	21.8	196.95	26.8	30.4	25.7	143.2	26.6
Up to 1040 per month	15.5	15.9	***	16.8	16.8	14.4	16.4	***	16.2	17	15.2	***	15.7
Up to 1380 per month	11.5	10.2		11	11.3	10.3	10.8		11.7	11.8	10.1		10.8
Up to 1730 per month	9.2	8.1		6.5	9.6	8.1	9		8.9	11.2	7.9		8.6
Up to 2080 per month	9	6.9		6.6	8.8	7.5	7.6		8	8.3	7.6		7.8
Higher than 2080 per month	17.8	11.6		9.5	15.3	14.6	14.5		14.2	13.3	14.6		14.3

Table 4 (continued)

	Subgroups by: sex			Subgroups by: age					Subgroups by: panel provider				Overall sample
	Male	Female	Chi^2	18 to 25	26 to 45	46 to 65	66+	Chi^2	MRFGR	Prolific	STV	Chi^2	
	%	%	Signif	%	%	%	%	Signif	%	%	%	Signif	%
I prefer not to answer	13.3	18.5		22.6	10.6	16.7	19.9		14.2	8.1	18.9		16.2
Employment:													
Full-time or self-employment	46.3	35.9	620.27	40	66.9	43.6	3.3	9775.65	44.5	45.3	37.6	391.65	40.5
Part-time or self-employment	8.7	19.4	***	16.3	16.2	18.1	5.9	***	13	20.7	14.2	***	14.7
Job-seeking	2.1	1.8		7.4	2.2	1.8	0		2.6	3.4	1.3		2
Long-term sick or disabled	5.5	7.3		4.1	6.1	10.4	0.9		6.2	3.4	7.3		6.5
Carer	1.3	2.1		1.2	1.5	2.4	1		1.8	1.2	1.8		1.7
Looking after the home and family	0.7	4.5		2.3	3.8	3.6	0.5		3.5	4.3	2.2		2.8
Retired	32	23.4		0.1	0.1	17.2	87.4		22.6	15.8	31.6		27.1
Student	1.7	2.8		24.5	1.2	0.2	0		3.5	3.9	1.5		2.4
Other	1	2		2.8	1.2	2	0.7		1.4	1.7	1.6		1.5
I prefer not to answer	0.7	0.8		1.3	0.9	0.8	0.2		0.9	0.4	0.8		0.7
The quality of your home in terms of repairs facilities and warmth:													
Good	68.7	66.9	8.88	55.8	62.1	68.2	77.3	273.35	62.5	64.3	70.8	99.25	67.6
Fair	25.8	28.1	*	35.5	31.6	26.5	20.3	***	31.5	29.8	24.6	***	27.2
Poor	5	4.6		7.2	6	5	2.2		5.3	5.5	4.4		4.8
I prefer not to answer	0.4	0.3		1.5	0.3	0.3	0.1		0.8	0.3	0.2		0.4
You are concerned about the safety of the neighbourhood you live in:													
Hardly ever	67.5	67.2	23.47	53.2	63.2	67.8	76.2	262.25	59.4	67.2	71.1	199.13	67.3
Some of the time	28	29.7	***	38.5	32.5	28.9	21.4	***	34.3	29.8	26.2	***	28.9
All of the time	4.2	2.8		6.9	3.9	3.1	2.1		5.5	3.1	2.5		3.4
I prefer not to answer	0.3	0.4		1.4	0.4	0.2	0.2		0.8	0	0.2		0.4

Table 5 Data quality checking

	Panel provider								
	MRFGR		Prolific		STV		Total		Chi^2
	No	%	No	%	No	%	No	%	Signif
Consistency SIPHER & SF-12v2 Physical health items									
Inconsistent	1,697	48.5	576	38.4	3,199	43.7	5,472	44.5	47.26
Consistent	1,800	51.5	924	61.6	4,114	56.3	6,838	55.5	***
Consistency SIPHER & SF-12v2 emotional health items									
Inconsistent	1,648	47.2	532	35.5	2,899	39.6	5,079	41.3	79.96
Consistent	1,842	52.8	966	64.5	4,423	60.4	7,231	58.7	***
Prefer not to say									
Did not choose "prefer not to say" at any point in the survey	2,748	77.8	1,307	87.1	5,334	72.4	9,389	75.7	157.02
Choose "prefer not to say" once or more times	786	22.2	194	12.9	2,032	27.6	3,012	24.3	***
Time to complete survey Median 13 min, IQR [8.592]									
Less than 6 min (5 percentile)	316	8.9	54	3.6	86	1.2	456	3.7	737.82
6–12 min (5–49 percentile)	1,700	48.1	932	62.1	3,084	41.9	5,716	46.1	***
13–74 min (50–95 percentile)	1,384	39.2	513	34.2	3,713	50.4	5,610	45.2	
74 min or more (top 5 percentile)	134	3.8	2	0.1	483	6.6	619	5.0	

Note:* $p<0.05$, ** $p<0.01$, *** $p<0.001$

Table 6 Equivalent income for the whole sample and for subsets of respondents with different SIPHER-7 profiles

	N	Mean	SD	Median	IQR
Equivalent income for the full sample	10,365	510.04	539.70	313.38	575.12
Subsample 1: best state for non-income SIPHER-7 items with full-time employment	698	1258.67	803.19	1099.91	1291.18
Subsample 2: best state for non-income SIPHER-7 items with retired	584	662.64	397.92	552.24	863.02
Subsample 3: best state for non-income SIPHER-7 items part-time employment	217	854.86	638.95	725.17	934.62
Subsample 4: some of the time lonely, full-time employment and the rest in the best state for non-income SIPHER-7 items	161	1002.29	686.74	794.21	1081.95

and binary sex as explanatory variables. Figure 3 shows the observed and predicted SF-12v2 MCS scores from the models, with the diagonal (complete agreement) indicated in red. Further analyses show that adding dummies to control for education, ethnicity, household size, personal income, dependent children or caring responsibility have little to modest effects on model performance (details available on request). Sensitivity analyses excluded respondents with poor data quality, such as those who answered the test–retest questions inconsistently or who completed the survey in less than 6 minutes or over 74 minutes (top and bottom 5th percentile). The final model is relatively robust, with no significant differences in the model coefficients between the sensitivity models and the final model (see Supplementary material).

Table 7 Summary statistics of instruments (total scores/utilities)

	N	Mean	SD	Median	IQR
Equivalent income (social perspective)	10,365	510.04	539.7	313.38	575.12
EQ-HWB-S tariff value	12,401	.75	.24	.83	.28
EQ-5D-5L tariff (Devlin et al,. 2018)	12,401	.82	.21	.88	.19
EQ-5D-VAS:	12,396	72.22	20.9	80	30
HUI3 tariff value	11,970	.71	.29	.79	.34
SF-6Dv1 tariff value	12,401	.71	.14	.66	.22
ICECAP-A tariff value	12,061	.78	.19	.85	.25
SF12v2 Physical Component Summary	12,401	49.75	10.61	52.77	14.09
SF12v2 Mental Component Summary	12,401	45.19	11.42	46.77	17.18
SWEMWBS total score	12,401	24.15	5.37	24	7
ONS-4 life satisfaction	12,401	6.64	2.33	7	3
ONS-4 worthwhile	12,401	6.75	2.46	7	4
ONS-4 happiness	12,401	6.63	2.5	7	4
ONS-4 anxiety	12,401	3.73	3.1	3	5

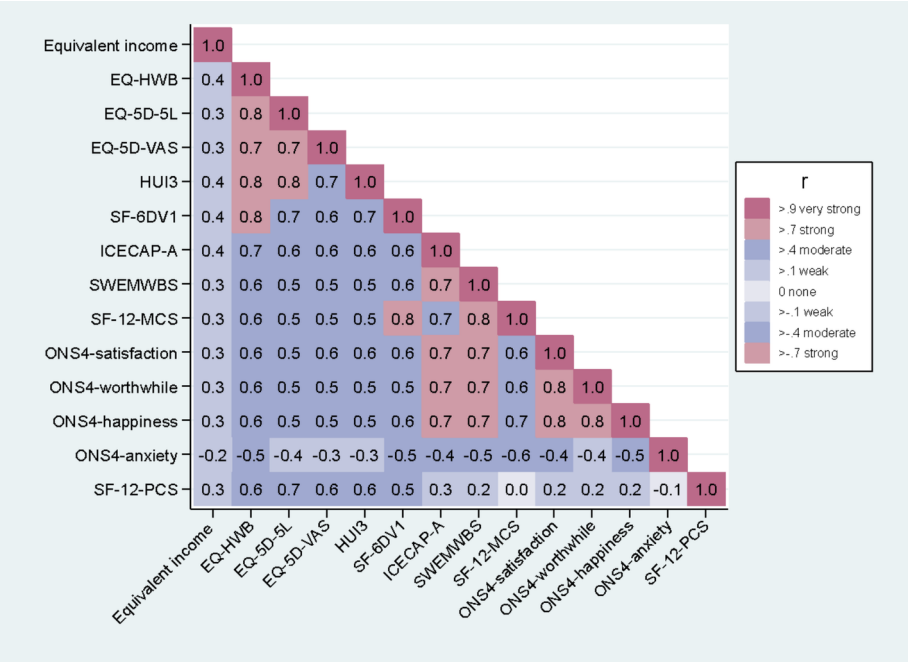

Fig. 1 Correlation coefficient matrix of the instruments

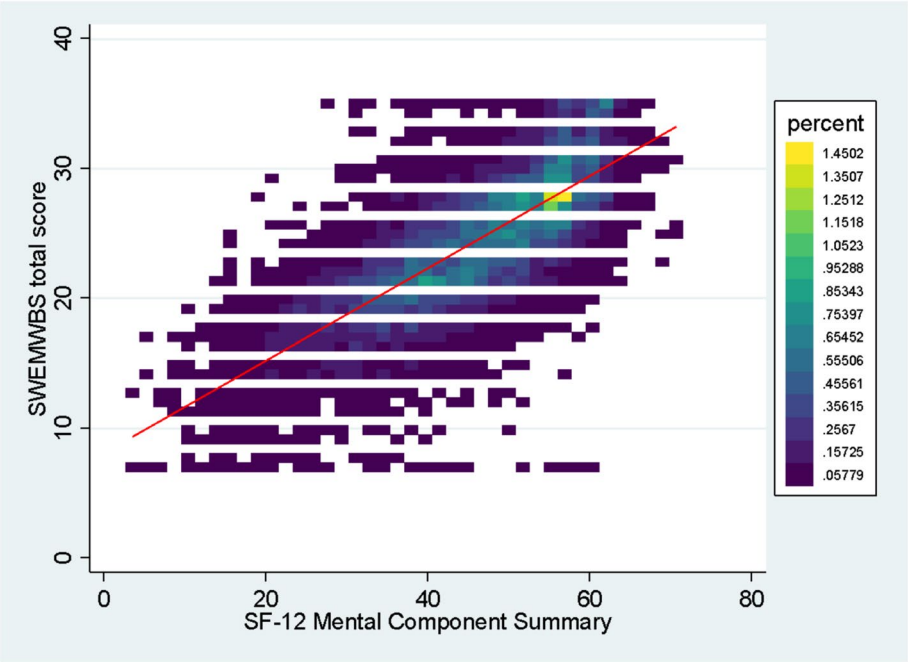

Fig. 2 Scatterplot of SWEMWBS total score vs SF-12v2MCS

4 Conclusion

This paper introduces a large-scale multi-instrument comparison dataset in health and wellbeing that includes a number of established health and wellbeing outcome measures, housing-related and work-related questions, alongside a series of socioeconomic variables, obtained online from over 12,000 members of the UK public aged 18+across three internet panels. The paper outlines the set-up of the survey, presents the key descriptive statistics, and reports on a basic mapping exercise between two established mental health outcome measures.

Existing MIC datasets include the SIPHER-HWMIC dataset, the UKHLS, the Health Survey for England (HSE), the Health Improvement and Patient Outcomes (HIPO) dataset, the South Yorkshire Cohort dataset (SYC65), the MIC dataset (Richardson et al., 2015) and the P-MIC dataset (Mukuria et al., 2016). Three of these datasets included members of the general public (SIPHER-HWMIC, UKHLS, HSE); the SYC65 and MIC datasets recruited both members of the general population and patients; the P-MIC dataset recruited children and oversample some with certain health conditions; and the HIPO dataset recruited patients recently discharged from hospital. All other datasets had large sample sizes (UKHLS over 50,000; HSE over 14,000; SIPHER-HWMIC over 12,000; MIC 9150; HIPO 6452; P-MIC 5000), with the exception of SYC65 (n=1749). All datasets were cross sectional except for UKHLS which is a panel survey and the HSE, which is a repeated cross-section survey. Most data were collected in online surveys, but some allowed paper-based data collection (UKHLS, HSE, SYC65, and HIPO). The MIC dataset contained the most number

Table 8 Modelled results of SF-12v2MCS and SWEMWBS items

	Model 1: all seven items as categorical	Model 2: collapse non-ordered levels	Model 3: as model 2, but add the 3-best dummy	Model 4: as model 3, including age and sex
Feeling optimistic:				
None of the time	0.000	0.000	0.000	0.000
	(.)	(.)	(.)	(.)
Rarely	1.399***	1.343***	1.335***	1.576***
	(0.317)	(0.318)	(0.318)	(0.312)
Some of the time	2.805***	2.773***	2.762***	3.331***
	(0.322)	(0.323)	(0.323)	(0.317)
Often	3.140***	-	-	-
	(0.346)			
All of the time	2.045***	-	-	-
	(0.448)			
Often/All of the time	-	2.927***	2.883***	3.990***
		(0.345)	(0.345)	(0.341)
Feeling useful:				
None of the time	0.000	0.000	0.000	0.000
	(.)	(.)	(.)	(.)
Rarely	0.922*	0.423	0.417	0.432
	(0.427)	(0.423)	(0.422)	(0.415)
Some of the time	2.771***	2.435***	2.433***	2.464***
	(0.431)	(0.427)	(0.426)	(0.419)
Often	4.135***	3.829***	3.821***	3.969***
	(0.447)	(0.443)	(0.443)	(0.436)
All of the time	4.457***	4.004***	3.750***	4.226***
	(0.503)	(0.498)	(0.503)	(0.494)
Feeling relaxed:				
None of the time	0.000	0.000	0.000	0.000
	(.)	(.)	(.)	(.)
Rarely	2.216***	1.959***	1.951***	2.000***
	(0.411)	(0.405)	(0.405)	(0.397)
Some of the time	6.444***	6.393***	6.391***	6.031***
	(0.419)	(0.413)	(0.413)	(0.404)
Often	10.318***	10.245***	10.224***	9.365***
	(0.444)	(0.438)	(0.438)	(0.430)
All of the time	12.093***	11.835***	11.678***	10.503***
	(0.526)	(0.519)	(0.521)	(0.513)
Dealing with problems well:				
None of the time	0.000	0.000	0.000	0.000
	(.)	(.)	(.)	(.)
Rarely	0.714	−0.021	−0.014	−0.094
	(0.514)	(0.489)	(0.489)	(0.480)
Some of the time	2.325***	2.211***	2.231***	1.944***
	(0.519)	(0.486)	(0.485)	(0.477)
Often	3.982***	3.915***	3.935***	3.543***
	(0.539)	(0.510)	(0.510)	(0.500)
All of the time	4.952***	4.933***	4.526***	4.047***
	(0.588)	(0.562)	(0.574)	(0.562)
Thinking clearly:				

Table 8 (continued)

	Model 1: all seven items as categorical	Model 2: collapse non-ordered levels	Model 3: as model 2, but add the 3-best dummy	Model 4: as model 3, including age and sex
None of the time/Rarely/Some of the time	-	0.000 (.)	0.000 (.)	0.000 (.)
None of the time	0.000 (.)	-	-	-
Rarely	−2.106*** (0.629)	-	-	-
Some of the time	−0.035 (0.636)	-	-	-
Often	3.231*** (0.650)	3.572*** (0.190)	3.591*** (0.190)	3.245*** (0.186)
All of the time	4.585*** (0.680)	5.359*** (0.274)	4.924*** (0.301)	4.242*** (0.295)
Feeling close to other people:				
None of the time	0.000 (.)	0.000 (.)	0.000 (.)	0.000 (.)
Rarely	2.054*** (0.403)	1.652*** (0.399)	1.656*** (0.399)	1.739*** (0.392)
Some of the time	3.804*** (0.399)	3.507*** (0.394)	3.515*** (0.394)	3.590*** (0.387)
Often	4.493*** (0.412)	4.199*** (0.408)	4.208*** (0.408)	4.352*** (0.401)
All of the time	4.837*** (0.450)	4.705*** (0.446)	4.405*** (0.454)	4.647*** (0.445)
Make my own mind up about things:				
None of the time	0.000 (.)	-	-	-
Rarely	−3.024*** (0.697)	-	-	-
Some of the time	−2.897*** (0.682)	-	-	-
Often	−1.163 (0.686)	-	-	-
All of the time	0.116 (0.693)	-	-	-
None/Rarely/Some of the time	-	0.000 (.)	0.000 (.)	0.000 (.)
Often/All of the time	-	2.158*** (0.185)	2.152*** (0.185)	1.765*** (0.182)
Best level:				
2 or less at best	-	-	0.000 (.)	0.000 (.)
3 or more at best	-	-	1.336*** (0.384)	0.989** (0.375)
Sex:				
Male	-	-	-	0.000 (.)

Table 8 (continued)

	Model 1: all seven items as categorical	Model 2: collapse non-ordered levels	Model 3: as model 2, but add the 3-best dummy	Model 4: as model 3, including age and sex
Female	-	-	-	−1.299*** (0.130)
Age group:				
18 to 25	-	-	-	0.000 (.)
26 to 45	-	-	-	1.134*** (0.255)
46 to 65	-	-	-	3.363*** (0.248)
66+	-	-	-	4.700*** (0.270)
Constant	25.027*** (0.631)	22.730*** (0.478)	22.745*** (0.478)	21.229*** (0.518)
Observations	12,401	12,401	12,401	12,357
R2	0.6098	0.6054	0.6058	0.6240
Adjusted-R2	0.6089	0.6047	0.6051	0.6232
Root-MSE	7.1401	7.1782	7.1750	6.9948

Note: Standard errors in parentheses; $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$

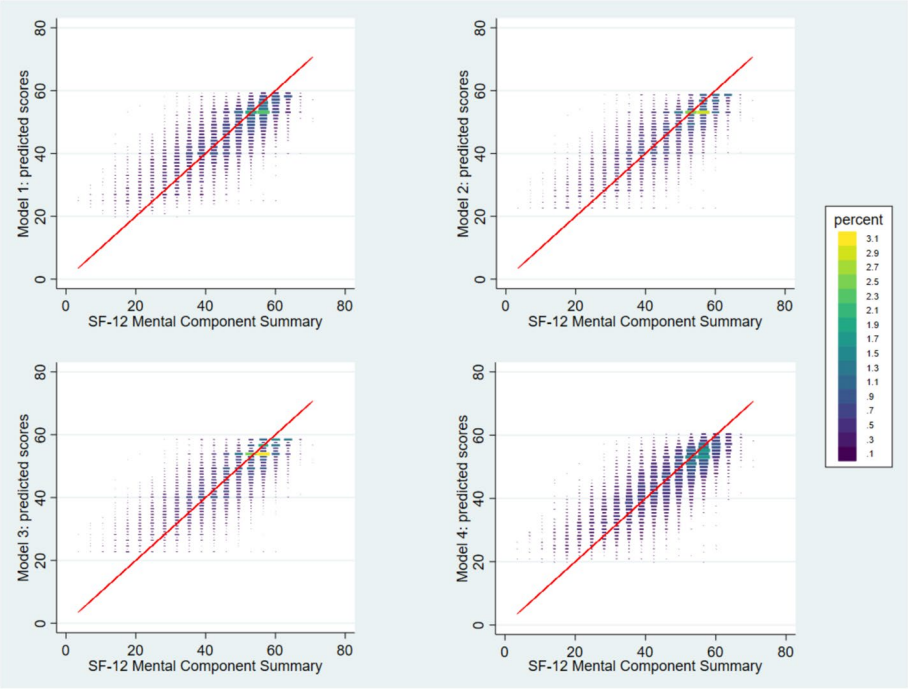

Fig. 3 Predicted values vs actual SF-12v2 MCS scores

(11) of outcome measures (personal well-being index (PWI), ONS-4, satisfaction with life survey (SWLS), EQ-5D-5L, AQoL-8D and AQoL-4D, HUI3, 15D, QWB-SA, SF-36, and ICECAP-A); followed by the SIPHER-HWMIC dataset with eight measures; the P-MIC included six outcome measures (EQ-5D-Y, CHU9D, PedsQL, HUI3, AQOL-6D, and PROMIS-25); the SYC6D included five outcome measures (EQ-5D-5L, ONS-4, WEMWBS, ICECAP-O and ASCOT); the HIPO dataset included four outcome measures (EQ-5D-5L, SF-12v2, ONS-4, SWB-VAS); the UKHLS included three outcome measures (SWEMWBS, GHQ-12 and SF-12v2); similarly, the HSE included three outcome measures (GHQ-12, WEMWBS, EQ-5D-3L).

Although our study was a large scale survey using a number of commonly used outcome measures, there are some limitations. First, this survey (as with the other existing MIC studies) was a cross-sectional study, and as such, the data do not allow for comparisons over time, or controlling for unobserved time-invariant individual heterogeneity. Second, the sample was broadly representative of the general public in terms of age and sex, but not other characteristics such as ethnicity, income, household size etc. This may limit the generalisability of findings. While it is possible to use population weights to adjust for any differences in these sociodemographic characteristics, as an online survey, the dataset excludes people who do not use the internet.

Third, there were differences in data quality across panels, with those belonging to the MRFGR panel more likely to speed through the survey and provide inconsistent responses when the same question is flipped. It is possible to conduct sensitivity analyses by setting data quality standards a priori to assess the robustness of results in future studies. For example, it is possible to exclude 'speeders' such as those who completed survey in under 6 minutes, or people who took longer (74 minutes and above) from the analysis and compare the results to assess the internal validity of the findings. Another form of sensitivity analysis is to exclude respondents who failed to consistently answer the repeated task question. The sensitivity analyses conducted in the mapping section demonstrated the overall robustness of the final model's internal validity, thereby increasing confidence in the reliability and accuracy of the results.

The purpose of this paper was to introduce the SIPHER- HWMIC data set, and as such the analyses conducted are primarily descriptive, with one example of mapping between SF-12v2 MCS and SWEMBS. Further studies can be conducted to map across different instruments, for example comparing EQ-5D-5L with other preference-based instruments. Also, psychometric analyses can be conducted to look for item response distribution and known-group validity. The survey contained items on good work and housing data, which can be used in further analyses, for instance to predict determinants of health. Moreover, subgroup analyses can be conducted to look at panel effects and differences between Scotland vs the rest of the UK. To conduct these potential analyses, the dataset is available from the UK Data Service (SN9458) for secondary use.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11205-025-03728-1.

Acknowledgements This research was conducted as part of the Systems Science in Public Health and Health Economics Research - SIPHER Consortium [https://sipher.ac.uk/] and we thank the whole team for valuable input that have informed this paper. The SIPHER Consortium was supported by the UK Prevention Research Partnership [https://mrc.ukri.org/research/initiatives/prevention-research/ukprp/] (Grant MR/S037578/2), which is funded by the British Heart Foundation, Cancer Research UK, Chief Scientist Office of the Scot-

tish Government Health and Social Care Directorates, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Health and Social Care Research and Development Division (Welsh Government), Medical Research Council, National Institute for Health Research, Natural Environment Research Council, Public Health Agency (Northern Ireland), The Health Foundation and Wellcome. We are grateful for the permissions given to use the following instruments: EQ-5D-5L, SF-12v2, ICECAP-A, EQ-HWB-S, and SWEMWBS. Views expressed by the authors in the publication do not necessarily reflect the views of EuroQol. Reproduction of Experimental EQ-HWB is not allowed. We would like to thank all survey respondents who participated. The dataset can be accessed from the UK Data Service (SN9458) [URL: https://doi.org/10.5255/UKDA-SN-9458-1].

Funding UK Prevention Research Partnership,MR/S037578/2, Petra Meier

Declarations

Ethical approval The survey was approved by the University of Sheffield's School of Health and Related Research Ethics Committee (ref: 049491).

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Ara, R., Rowen, D., & Mukuria, C. (2017). The use of mapping to estimate health state utility values. *PharmacoEconomics, 35*, 57–66.

Brooks, R., & the EuroQol Group. (1996). EuroQol: The current state of play. *Health Policy, 37*(1), 53–72. https://doi.org/10.1016/0168-8510(96)00822-6

Brazier, J. E., & Roberts, J. (2004). The estimation of a preference-based measure of health from the SF-12. *Medical Care*. https://doi.org/10.1097/01.mlr.0000135827.18610.0d

Devlin, N. J., Shah, K. K., Feng, Y., Mulhern, B., & van Hout, B. (2018). Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health Economics, 27*, 7–22. https://doi.org/10.1002/hec.3564

Feeny, D., Furlong, W., Torrance, G. W., Goldsmith, C. H., Zhu, Z., DePauw, S., Denton, M., & Boyle, M. (2002). Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Medical Care*. https://doi.org/10.1097/00005650-200202000-00006

Fleurbaey, M. (2006). Health, Equity and Social Welfare. *Annales d'_Economie et de Statistique*, (83/84): 21–59.

Fleurbaey, M. and Blanchet, D. (2013). *Beyond GDP: Measuring welfare and assessing sustainability*. Oxford University Press.

Fleurbaey, M. (2005). Health, wealth, and fairness. *Journal of Public Economic Theory, 7*(2), 253–284.

Fleurbaey, M. (2009). Beyond GDP: The quest for a measure of social welfare. *Journal of Economic Literature, 47*(4), 1029–1075.

Fleurbaey, M., & Gaulier, G. (2009). International comparisons of living standards by equivalent incomes. *Scandinavian Journal of Economics, 111*(3), 597–624.

Flynn, T. N., Huynh, E., Peters, T. J., Al-Janabi, H., Clemens, S., Moody, A., & Coast, J. (2015). Scoring the ICECAP-A capability instrument: Estimation of a UK general population tariff. *Health Economics, 24*(3), 258–269.

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

Hernández, M. H., Wailoo, A., Pudney, S., Gray, L., & Manca, A. (2020). Mapping clinical outcomes to generic preference-based outcome measures: Development and comparison of methods. *Health Technology Assessment (Winchester, England), 24*(34), 1.

Hernández-Alava, M., S. Pudney, and A. J. Wailoo. (2020). Estimating the relationship between EQ-5D-5L and EQ-5D-3L: Results from an English Population Study. National Institute for Health Research (NIHR) Policy Research Unit in Economic Methods of valuation in Health & Social Care Interventions. Universities of Sheffield and York. Report no. 063. http://www.eepru.org.uk/wp-content/uploads/2020/10/eq5d-5l-final-report-30-9-20.pdf.

Jones, R., Mulhern, B., McGregor, K., Yip, S., O'Loughlin, R., Devlin, N., Hiscock, H., Dalziel, K., Quality Of Life in Kids: Key Evidence to Strengthen Decisions in Australia (QUOKKA) Project Team. (2021). Psychometric performance of HRQoL measures: An Australian Paediatric Multi-Instrument Comparison Study Protocol (P-MIC). *Children, 8*, 714. https://doi.org/10.3390/children8080714

Meier, P., et al. (2019). The SIPHER consortium: Introducing the new UK hub for systems science in public health and health economic research. *Wellcome Open Research*. https://doi.org/10.12688/wellcomeopenres.15534.1

Mukuria, C., Rowen, D., Peasgood, T. and Brazier, J., (2016). *An empirical comparison of well-being measures used in UK. National Institute for Health Research (NIHR) Policy Research Unit in Economic Methods of valuation in Health & Social Care Interventions*. Universities of Sheffield and York. Report no. 027.

Mukuria, C., Peasgood, T., McDool, E., Norman, R., Rowen, D., & Brazier, J. (2023). Valuing the EQ health and wellbeing short (EQ-HWB-S) using time trade-off and a discrete choice experiment: A feasibility study. *Value in Health*. https://doi.org/10.1016/j.jval.2023.02.008

Office for National Statistics. (2018). Personal well-being user guidance. [Accessed 22 December 2022]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/methodologies/personalwellbeingsurveyuserguide

Office for National Statistics, (2022). Estimates of the population for the UK, England, Wales, Scotland and Northern Ireland. [Accessed 28 January 2023]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland

Peasgood, T., Brazier, J.E., Mukuria, C. and Rowen, D. (2014). A conceptual comparison of well-being measures used in the UK. National Institute for Health Research (NIHR) Policy Research Unit in Economic Methods of valuation in Health & Social Care Interventions. Universities of Sheffield and York. Report no. 026.

Petrou, S., Rivero-Arias, O., Dakin, H., Longworth, L., Oppe, M., Froud, R., & Gray, A. (2015). Preferred reporting items for studies mapping onto preference-based outcome measures: The MAPS statement. *Medical Decision Making, 35*(6), NP1–NP8.

Richardson, J., Khan, M. A., Iezzi, A., & Maxwell, A. (2015). Comparing and explaining differences in the magnitude, content, and sensitivity of utilities predicted by the EQ-5D, SF-6D, HUI 3, 15D, QWB, and AQoL-8D multiattribute utility instruments. *Medical Decision Making, 35*(3), 276–291.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia and AnalgesIa, 126*(5), 1763–1768.

Tsuchiya, A. and Wu, C. on behalf of the SIPHER Consortium (2021). SIPHER-7: a seven-indicator outcome measure to capture wellbeing for economic evaluation, Working paper, SIPHER Research Paper Series 1. URL: https://sipher.ac.uk/wp-content/uploads/2021/10/Sipher-7-report.pdf

Tsuchiya, A. Wickramasekera, N and Field, B. on behalf of the SIPHER Consortium (2022). Review of SIPHER-7 wellbeing indicators, Working paper, SIPHER Research Paper Series. URL: https://sipher.ac.uk/wp-content/uploads/2022/12/Review-of-Sipher-7-wellbeing-indicators.pdf

University of Warwick. (2011). Collect, score, analyse and interpret WEMWBS. URL: https://warwick.ac.uk/fac/sci/med/research/platform/wemwbs/using/howto/wemwbs_population_norms_in_health_survey_for_england_data_2011.pdf

University of Essex, Institute for Social and Economic Research. (2020). Understanding Society: Waves 1-10, 2009-2019 and Harmonised BHPS: Waves 1-18, 1991-2009. [data collection]. 13th Edition. UK Data Service. SN: 6614, 10.5255/UKDA-SN-6614-14.

Wailoo, A. J., Hernandez-Alava, M., Manca, A., Mejia, A., Ray, J., Crawford, B., & Busschbach, J. (2017). Mapping to estimate health-state utility from non–preference-based outcome measures: An ISPOR good practices for outcomes research task force report. *Value in Health, 20*(1), 18–27.

Ware, J.E., Kosinski, M., Turner-Bowker, D.M. and Gandek, B. (2002). *How to score version 2 of the SF-12 health survey*. Lincoln, RI: QualityMetric Incorporated.

Wickramasekera, N., Field, A., Thu Ta, A., and Tsuchiya, A. (forthcoming). Using discrete choice experiments (DCEs) to compare social and personal preferences for health. Medical Decision Making.
Wolowacz, S. (2017). New ISPOR recommendations-mapping methods for estimation of health state utility. *Value in Health, 20*(1), 28–29.