# Accounts from developers of generic health state utility instruments explain why they produce different QALYs: a qualitative study

Pickles, Kristen; Lancsar, Emily; Seymour, Janelle; Parkin, David; Donaldson, Cam; M. Carter, Stacy

*Published in:*
Social Science and Medicine

*DOI:*
10.1016/j.socscimed.2019.112560

*Publication date:*
2019

*Document Version*
Publisher's PDF, also known as Version of record

Social Science & Medicine 240 (2019) 112560

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

# Accounts from developers of generic health state utility instruments explain why they produce different QALYs: A qualitative study

Kristen Pickles[a], Emily Lancsar[b,*], Janelle Seymour[c], David Parkin[d], Cam Donaldson[e], Stacy M. Carter[f]

[a] Sydney School of Public Health, Edward Ford Building A27, The University of Sydney, NSW 2006, Australia
[b] Department of Health Services Research and Policy, Research School of Population Health, The Australian National University, 63 Eggleston Road, Acton, ACT 2601, Australia
[c] Centre for Evaluation and Research, Department of Health and Human Services, 50 Lonsdale Street, Melbourne, VIC 3000, Australia
[d] City University, London and Office of Health Economics, London, UK
[e] Yunus Centre for Social Business and Health, Glasgow Caledonian University, Cowcaddens Road, Glasgow, G40BA, UK
[f] Australian Centre for Health Engagement, Evidence and Values (ACHEEV), School of Health and Society, Faculty of Social Sciences, Building 15 Room 240, University of Wollongong, NSW 2522, Australia

ARTICLE INFO

*Keywords:*
Australia
North America
Europe
Preference weighted quality of life instruments
Health state utility instruments
Health Utilities Index (HUI)
EQ-5D
Short Form 6D (SF-6D)

ABSTRACT

*Purpose and setting:* Despite the label "generic" health state utility instruments (HSUIs), empirical evidence shows that different HSUIs generate different estimates of Health-Related Quality of Life (HRQoL) in the same person. Once a HSUI is used to generate a QALY, the difference between HSUIs is often ignored, and decision-makers act as if 'a QALY is a QALY is a QALY'. Complementing evidence that different generic HSUIs produce different empirical values, this study addresses an important gap by exploring how HSUIs differ, and processes that produced this difference. 15 developers of six generic HSUIs used for estimating the QOL component of QALYs: Quality of Well-Being (QWB) scale; 15 Dimension instrument (15D); Health Utilities Index (HUI); EuroQol EQ-5D; Short Form-6 Dimension (SF-6D), and the Assessment of Quality of Life (AQoL) were interviewed in 2012–2013.

*Principal findings:* We identified key factors involved in shaping each instrument, and the rationale for similarities and differences across measures. While HSUIs have a common purpose, they are distinctly discrete constructs. Developers recalled complex developmental processes, grounded in unique histories, and these backgrounds help to explain different pathways taken at key decision points during the HSUI development. The basis for the HSUIs was commonly not equivalent conceptually: differently valued concepts and goals drove instrument design and development, according to each HSUI's defined purpose. Developers drew from different sources of knowledge to develop their measure depending on their conceptualisation of HRQoL.

*Major conclusions/contribution to knowledge:* We generated and analysed first-hand accounts of the development of the HSUIs to provide insight, beyond face value, about how and why such instruments differ. Findings enhance our understanding of why the six instruments developed the way they did, from the perspective of key developers of those instruments. Importantly, we provide additional, original explanation for why a QALY is not a QALY is not a QALY.

Quality Adjusted Life Years (QALYs) are the dominant measure of health benefit used in economic evaluation to inform health care resource allocation decisions. QALYs account for both length of life and health related quality of life (HRQoL) in a single index. Over the past four decades, a small number of generic preference-weighted health state utility instruments (HSUIs; also referred to as multi-dimension utility instruments) have been developed to measure HRQoL for use in QALYs.

Health state utility measurement comprises two main elements: (a) a health state classification system: defining and describing a set of health states of interest, usually presented as a standardised questionnaire, and (b) valuation of those health states to generate the HRQoL weights used to generate QALYs.

Six generic HSUIs are used most widely for estimating the quality of

* Corresponding author.
*E-mail address:* Emily.Lancsar@anu.edu.au (E. Lancsar).

https://doi.org/10.1016/j.socscimed.2019.112560
Received 22 January 2019; Received in revised form 16 September 2019; Accepted 17 September 2019
Available online 19 September 2019

life component of QALYs (listed in chronological order of development): the Quality of Well-Being (QWB) scale; 15 Dimension instrument (15D); Health Utilities Index (HUI); EuroQol EQ-5D; Short Form-6 Dimension (SF-6D), and the Assessment of Quality of Life (AQoL) (Drummond et al., 2005). All are examined in this study and described in detail in Table 1 in the next section. EQ-5D is the most extensively used HSUI worldwide (Richardson et al., 2014; Brazier et al., 2017a,b). In practice, funding bodies and academics use a variety of HSUIs to inform their decisions. Although a small number of health technology assessment (HTA) bodies have specified which measure should be used—for example, the National Institute for Health and Care Excellence (NICE) in the UK has specified the EQ-5D as a preferred measure—typically HTA bodies will accept measures from any of the above HSUIs (Longworth and Longson, 2008).

The claimed theoretical advantage of generic HSUIs is their ability to produce values comparable across all interventions and diseases, thereby producing a common currency for economic evaluation, including for use in HTA (Finch et al., 2018). However although the six instruments are termed generic and used in economic evaluation as if they are homogenous, they differ in their constructs, including in content and size of their descriptive systems; valuation methods; and populations used to value the health states (Brazier et al., 2017a,b). Further, as demonstrated in head-to-head empirical comparisons, they also generate different utility values (Fryback et al., 2010; Hawthorne et al., 2001).

Many health economists (and others) understand that different HSUIs will: 1) generate different empirical estimations of HRQoL in the same person; 2) thus produce a different resulting number of QALYs; 3) generate, therefore, different ratios in cost utility analyses of the same intervention; and 4) potentially, result in different funding decisions, depending on which HSUI is used (Richardson et al., 2016; Brazier et al., 2017a,b). In practice, however, this difference is often ignored, and decision-makers act as if 'a QALY is a QALY is a QALY' regardless of the instrument.

In this study we go beyond the recognised fact that the instruments produce different empirical values, to explore and explain how and why HSUIs differ. To do this, we generated and analysed unique data focusing on first-hand accounts of the development of the six HSUIs to provide insight about the development process, from the perspective of key developers of those instruments. In particular, this study aims to:

1. Explore factors that influenced the original development of six HSUIs used to calculate QALYs
2. Explain why HSUIs differ in their production of QALYs, from the perspective of 15 developers of six HSUIs.

Some instrument developers have published accounts of the development of some individual instruments (Torrance et al., 1982; Kaplan and Anderson, 1988; Sintonen and Pekurinen, 1989; Brooks and Group, 1996; Torrance et al., 1996; Hawthorne et al., 1997; Brazier et al., 2002; Devlin and Brooks, 2017). Informative analyses of published literature (e.g (Richardson et al., 2016; Brazier et al., 2017a,b; Finch et al., 2018) compared the construction and validity base of existing instruments. Some accounts of the development of particular HSUIs refer to the use of qualitative methods to help generate descriptive systems (e.g. (Kaplan et al., 1997; Richardson et al., 2012) but standalone qualitative papers linked to particular HSUIs are generally not published, with some exceptions (e.g. (Herdman et al., 2011; Keeley et al., 2013; Stevens, 2017).

This current study is the first to explore development across instruments as explained by developers from all 6 generic HSUIs. This study is also the first to take a qualitative approach across a range of developers allowing a depth and nuance in analysis including a focus on the meaning and social value underlying HSUIs, which in our view was not able to be achieved in previous investigations of these measures. Understanding each instrument's origins and how and why key decisions were made throughout their development is expected to help prospective users to choose which HSUI to use in different contexts, researchers and policy makers to better understand the data that result from their use, and those considering developing new HSUIs to learn from existing examples of instrument development.

The literature has expanded beyond generic measures to condition-specific measures (e.g., instruments focusing on specific areas outside of health (e.g. ASCOT for measurement of social outcomes, excluding health status with a broader focus on quality of life) (Netten et al., 2012), and the emerging literature on the generation of quality of life measures more generally, for example including capability measures such as the ICECAP (Coast et al., 2008) and OxCAP (Lorgelly et al., 2008; Simon et al., 2013; Coast et al., 2015) measures, and the idea of 'super QALYs' such as the well-being adjusted life-year: WELBY, potentially for use across sectors such as social care and public health (Brazier and Tsuchiya, 2015).

Because our study illuminates the detail of similarities and differences in processes of instrument development, we expect that our findings are likely to be useful to these efforts also. In the spirit of accessibility, including to students, in the next section we provide an overview of the different measures and concepts before moving on to describe our methods and results. We conclude with a discussion of our findings, implications and areas for future research.

## 1. Background

QALYs provide a measure of health gain that incorporate both quantity and quality of life. In particular, QALYs are calculated as life expectancy multiplied by the HRQoL experienced in those years. We note that while 'Quality of Life' encompasses a broad range of factors that shape an individual's life, including non-health aspects, the concept of HRQoL has been defined in a number of ways. For example, Brazier et al. review four definitions used in the literature including: in relation to an individual's functioning and subjective wellbeing; factors that are part of an individual's health; aspects of quality of life affected by health – e.g. by the presence or absence of disease; and the value of health states, or utilities, which can be used to calculate QALYs (Torrance, 1987; Ebrahim, 1995; Weinstein et al., 1996; Killewo et al., 2010; Karimi and Brazier, 2016). We use the latter in this paper. The HRQoL index used to "quality adjust" life years is measured on a cardinal scale from 0 to 1 where 0 is dead and 1 is full health. So, for example, a person expected to live for a further 10 years in a health-related quality of life of 0.8 has 8 QALYs.

HSUIs are used to generate the HRQoL values used to quality adjust life years to generate QALYs. They focus on the utility associated with health states alone, rather than the broader economic concept of utility (Bleichrodt and Quiggin, 2013). Health states are described by a broad range of dimensions to provide response options that enable respondents to accurately describe their current health state. Examples include psychological wellbeing, pain, ability, and symptoms. They should ideally be based on direct patient experience, though many use clinicians as proxies backed by peer-reviewed literature (Tolley, 2009). HSUIs differ considerably in the content and size of their descriptive systems, valuation method applied, population used to value the health states, and scoring algorithms (See Table 1 supplementary and (Brazier et al., 2017a,b) for a detailed summary and comparison of the descriptive systems).

For example, for the six HSUIs included in this study, health states are described in terms of capacity (e.g. HUI), function (EQ-5D), or actual behaviour and performance (e.g. QWB), and the coverage of symptoms, mental health, and social health (including consequences for usual activities, work, and relationships) also varies (Brazier et al., 2017a,b). Only two dimensions – mobility and pain - are covered by all instruments.

Several "preference elicitation techniques" are used to value health states in HSUIs, including visual analogue scales (VAS); time trade off

(TTO), standard gamble (SG), and discrete choice experiments (DCEs) (Table 2 supplementary).

The VAS is best described as a psychometric response scale, with two anchors representing 'best imaginable' and 'worst imaginable' health. In contrast, TTO and SG derive weights by asking participants to choose between alternative scenarios (e.g. living longer in an impaired health state or living in full health for a shorter period; or remaining in a certain impaired health state or taking a gamble of being in full health or risking death) (Drummond et al., 2005). In recent times, discrete choice methods have also been used to value health states. The practicality of each method depends on its acceptability to respondents (including length and complexity of the task) and its ability to keep the respondents' interest (Brazier et al., 2017a,b). Another factor is whether people can actually understand the concepts that the valuation methods use. For example, the SG assumes people understand probabilities well and can handle probability calculations; TTO assumes people can handle implied trade-offs between length and quality of life. The choice of method matters because differences in their theoretical grounding and valuation approach can lead to differences in utility estimates for the same health states.

## 2. Methods

### 2.1. Ethics approval

Ethics approval was obtained for the study (Monash University Human Research Ethics Committee (MUHREC): project number CF12/2236–2012001201).

### 2.2. Approach

We chose a qualitative approach because this best allowed us to answer our research questions. We were genuinely uncertain what developers would consider to be important in their explanations of instrument development: a qualitative approach allowed us to learn more by remaining flexible in data collection and analysis. A qualitative approach also allowed for a more nuanced investigation than other more quantitative methods would have allowed.

### 2.3. Participant selection and recruitment

Key researchers responsible for the development and modification of the six instruments were identified from the literature. We employed a purposive sampling strategy (Patton, 2002; Given, 2008), with the goal of recruiting more than one informant per instrument. We did not approach all developers. Instead we identified potential informants based on when they had been involved in instrument development (e.g. whether they were originators of the instrument, or whether they joined the development team later). We made sure that at least one of the developers per instrument was one of the original developers and aimed to recruit participants from initial and subsequent stages of development. For EQ-5D, which had a larger development group than other instruments, we recruited more developers. Potential participants were approached via email. In accordance with the study's ethics approval, participants were provided with a participant information statement which explained what was involved in taking part in the study including that all results would be presented in a de-identified form with participant names removed. Participants were told that by replying to the email to be part of the study and arranging a time for interview they provided their consent.

### 2.4. Data collection

A pilot study (Wilson, 2010) provided important insights that guided the design of the main study and a draft topic guide.

We chose semi-structured individual interviews to give individual developers maximum freedom to explain their own perspective on their instrument; we interviewed by telephone because informants were located all over the world. The participants were interviewed using a topic guide, including questions on the instrument's background; aims; development criteria; descriptive system; preference weights; instrument performance; comparisons with other HSUIs; impact, and future of the instrument and the field more broadly. Interviews were conducted between November 2012 and August 2013. The average interview length was 54 min, ranging from 27 to 89 min.

One researcher (JS) conducted the interviews and recorded her reflections about each interview in field notes. As a health economist, the interviewer had a shared understanding of terminology used by participants. JS was part of the international health economics community at the time of data collection; she knew some participants on a professional basis only (e.g. via occasional contact at conferences and seminars). Interviews were audiotaped and transcribed verbatim. The field notes were discussed by researchers (JS and EL) following each interview.

### 2.5. Data analysis

Although the study design did not lend itself to theoretical sampling, we used Charmaz's iteration of grounded theory as an approach to analyse the qualitative data (Charmaz, 2006). Grounded theory analysis methods are relatively prescriptive: analysis proceeds in stages, beginning with line by line coding. This technique forces the analyst to go beyond surface impressions or themes, paying attention to detail and action. SMC conducted an initial analysis of three interviews. Early line by line codes included: quantifying the health of nations, valuing comprehensiveness, evolving, compromising, and allocating resources. Memo-writing is another key analytic method for grounded theory: memos are written about cases (individual informants: these memos capture what the analyst has learned from this interview) and concepts (these memos allow cross-case comparison about key concepts). SMC also wrote detailed memos about the first three interviews as well as a topline conceptual memo which suggested some directions for the overall analysis, including that it should code for making sense of the instrument (including the purpose of the instrument, and comparing instruments or defending one's own instrument), having goals (what goals people said they had and what these were), developing the instrument (coding for the nature of the development process) and valuing different characteristics (to capture differences in what developers took to be a 'good' instrument.

This broad direction for analysis was taken up by KP, who with support from SMC in consultation with EL continued to use line by line coding and constant comparison (a grounded theory method of comparing data across cases) to systematically sort, compare and categorise data. The overall aim was to identify similarities, differences, relationships and patterns so as to draw conclusions across all of the interviews (Green and Britten, 1998). KP read and coded transcripts of all interviews, developing and further refining the analytic categories, which included 'defending my instrument', 'having goals', and 'trading off'. KP continued to write memos, with a particular focus on how insights from each new interview reinforced or altered the categories and concepts initiated by SMC. SMC, EL, and KP met frequently throughout the analysis to discuss and develop emerging concepts, categories, and their relationships, and to ensure interpretive validity of preliminary findings. KP and SMC were guided by EL, DP, JS and CD's broader understanding of the relevant literature. The subheadings in our results section reflect the central categories under which our coding was eventually organised, and the text of the paper reflects the thinking recorded in our analytic memos and discussions.

Illustrative quotes are presented in the following section. Information provided at the end of each quotation includes an acronym for the HSUI that the developer worked on, 'p' denoting 'participant', and then numbering of the interviewees in the order in which they were

interviewed (e.g. EQP1, HUIP2).

## 3. Results

Of the 19 people invited to participate, 15 were interviewed. One declined due to retirement, one did not agree, and one did not respond after three reminders. Another agreed to be interviewed, but later was unavailable due to illness. In many instances all of the developers of a particular instrument were interviewed. The final sample contained two participants for each instrument except EQ-5D which had five participants to reflect the much larger development group for that instrument. Informants were from a range of disciplines with an interest in instrument development, including health economics, medicine, psychometrics, and health service research. As planned, we interviewed at least one participant involved in the original development of each instrument; other participants were involved in subsequent instrument development. Interviewees were based in North America, Europe and Australia and include the majority of global HSUI developers.

We found that while the six HSUIs are being used for the same common purpose – to deliver data for the construction of QALYs for economic evaluations – and are ostensibly the same 'type' of instrument, in practice they construct HRQoL in heterogeneous ways. The developer's accounts suggest a range of differences between the instruments. This included differences in: 1) sources of knowledge underpinning conceptualisations of health and wellness;; 2) development (purpose, goals, processes); 3) central qualities that were prioritised throughout instrument development; 4) selecting valuation techniques; and 5) contextual factors, including histories, collaborations, funders and shared values. These contrasts help to explain why different pathways were taken at key decision points during the HSUIs' development, resulting in quite different instruments. Supplementary Tables 3–5 provide some detail about developers' perceptions of their instrument and its development; these tables are explained in detail in subsequent sections.

## 4. Developers drew on different knowledge sources to conceptualise HRQoL differently

All developers aimed to create an instrument that measured HRQoL. However, this central variable was conceptualised differently by different teams. HRQoL is, in the words of one developer, a 'complex latent variable', difficult to describe, and developers *'don't even agree on what the latent variable is'* (QWBP2).

> These are all different concepts of what it is we think we want to measure. So it's a bit … I suppose we might call almost normative element first of all which is about what is it policy makers want to measure (SFP2).

> There is no accepted theory of HRQoL or health that would tell us what dimensions to include … by the WHO and things like that, what the instrument should contain conceptually … how to operationalise this problem and all the instruments work in different ways (15DP1).

The conception of each instrument was connected to an explicit purpose, formulated by individual HSUI developers, teams of developers, and/or funders (Table 3 supplementary).

The design and development of each instrument generally aligned with its ostensible purpose, and these were diverse. Some purposes were instrumental: focused on what the instrument would allow the user to do (monitor populations, allocate resources), others institutional: connected to health care system or policy goals (productivity comparisons, cost-effectiveness), or epistemic: relating to intrinsic value of the task for the developer (solving methodological puzzles, quantifying health). Some commitments shifted as the development progressed (e.g. initially a research task, eventually developed for use in population health surveys), and these only sometimes formally articulated.

To conceptualise HRQoL, developers relied on widely varying sources of knowledge, from authoritative lists through to doing original research from the ground up. Their choices in this regard tended to be consistent with their purpose for their instrument (Table 3). For instance, some drew from disciplinary or technical sources: QWB developers were guided by 'objective' medically oriented definitions from medical textbooks and clinical specialists, *'looking at how clinicians and epidemiologists asked about health'* (QWBP1), with a focus on symptoms. Some utilised health states described in existing instruments or datasets: three HSUIs began with a broad theory of QOL derived from the 1948 World Health Organisation (WHO) definition of health. Others were guided by exploratory work rather than drawing from existing conceptualisations, looking at *'the sorts of things people thought were important, from their health perspective, what things they value'* (15DP2)*; 'the attributes … that folks said were the most important'* (HUIP1), and undertook extensive population surveys. 15D developers drew from health policy documents to find out how health was conceptualised (e.g. ability to work). Each essentially constructed a unique classification system, reflecting the complexity and nuance of defining and describing HRQoL, the focus of measurement.

We also identified considerable variation in the processes by which health states were subsequently constructed and validated, both more or less formally, including via expert consensus between developers (*'thrashing it out and finally agreeing'* (EQP4)), empirical engagement (e.g. with clinicians, psychiatrists, patient groups), or using extensive statistical analyses. These technical and conceptual commitments, once secured, tended to be handed down in the culture of an instrument-development team, whether or not individual developers had a strong commitment to them.

## 5. Purpose, central concepts and development processes were connected

Although a range of commitments and circumstances drove development of each HSUI, we were able to observe a central defining concept at the heart of each HSUI's development (Table 4 supplementary) . This was our interpretation of the most central ideas that developers said guided the work, and which they used to explain decisions and actions taken. In doing this we are not suggesting that each developer cared about only one thing: rather, we sought to draw out the most important explanation for why things went the way they did for each HSUI. These concepts interacted to some extent with the stated purpose for the instrument, as discussed above.

### *5.1. An important contrast: Grand goals and purposeful process, or modest goals and opportunistic process*

Some developers recalled broad and ambitious goals for their HSUI, seemingly driven by public health bureaucracies and public policy. We observed that these developers referred to, and valued, doing a 'proper' job, that is, progressing in a serious way, with intention, systematically. Other developers described modest, discrete goals for their HSUI - to fill a perceived need, to incorporate methodological advancements, or to pursue academic interests. We observed that these developers described an informal, opportunistic process, with somewhat unexpected outcomes. Use of EQ-5D and HUI, for example, went greatly beyond their intended application and 'accidently' became stand-alone HSUIs. Instruments are classified using this distinction in Column 2, Table 4, and we provide examples in the section following. We note that this contrast was not absolute: there was a small amount of overlap between modest and opportunistic goals and grand and purposeful goals for some instruments (e.g. SF-6D). However, overall, the contrast was strong and was an important way of distinguishing different HSUI development processes.

HSUIs with grand goals progressed according to clearly defined institutional objectives and were influenced by the priorities of the respective health care systems. They were mostly large, funded projects (e.g. funded by government) with specific requirements evident from the outset. Scientific foundations, objectivity in design, and/or following the prescriptions of psychometric theory were of primary importance for developers of these HSUIs, which included AQoL, 15D, and QWB. The central driving concepts of instruments with a population-level focus tended to correspond with high expectations for the HSUI itself and the endpoints it could achieve (Table 4). For example, QWBP1's focus for the QWB was *Solving grand problems* of measurement and monitoring of population health, to demonstrate its superiority as a HSUI. 15DP1's account similarly indicated a mission of *Developing the best* (15DP1) instrument on the market, claiming that existing HSUIs were of poor quality. The AQoL was developed in response to perceived inadequacies of existing measures but with the core aim of *Starting from scratch, correctly*: AQoLP1 described his intention to change habitual approaches to HSUI development by setting a new gold standard incorporating methodological advancements and following prescriptions of correct psychometric theory. Developers of these instruments mostly described the impact of their HSUI in the context of national policy impact, clinical trial relevance, incorporation in population health surveys, and widespread use in economic evaluation (see Instrument Impact column in Table 4).

HSUIs with more modest stated purposes began as small-scale, personal interest projects. They had individually defined goals, generally confined to the research task at hand; for example, to find a common set of questions, to modify an existing health state measure, or to construct a meaningful measure for critical populations. They then 'accidently' became stand-alone HSUIs or progressed to a scale that *wasn't really fully anticipated* (HUIP2). Exploratory work and consensus-based decisions were central to the accounts of those developers with more modest goals, which included EQ-5D, HUI, and SF-6D. Their accounts suggested a focus on the development process, rather than achieving a specific endpoint. For example, EQP1 described an exploratory intellectual project of *Joining forces* (column 3) underlying the development of the EQ-5D, a joining of researchers across different disciplines and different countries. Most decision points for the EQ-5D were described as a collaborative process involving international interdisciplinary teams joining to solve a problem (how to measure health): *'it was not really a group to develop something; it was just a common interest to see what we could do together'* (EQP1). A participant involved in the development of the SF-6D, which originated from a doctoral thesis, described development meetings where *'it was literally a case where we would sit in Boston having a meeting and we'd go, "Well I don't know, I'm not sure if you really need that dimension" … she'd go away (and) a team of researchers somewhere in the building, just running loads of factor analysis and Rasch analysis, to see how we could reduce the instrument down and it was quite fun … It was really amazing* (SFP1).

Developers of (initially) small-scale HSUIs were also particularly attentive to capturing and reflecting social preferences. For example, the HUI was developed in the context of illness and *Reflecting lived experience* was critical to its developers; HUIP1 reported working empirically with regular people close to illness experiences to ensure descriptive items captured lived conditions. The impact of modest-goal HSUIs was framed in terms of their practical use (e.g. NICE recommended, practice evaluation), broad acceptability (straightforward and easy to use), contributing to awareness and debates about health valuation (i.e. it's 'approachable'), number of translations internationally, generation of revenue for research purposes, or with reference to intrinsic value, including helping decision makers (Table 4 Instrument Impact column).

Developers who referred to having followed a 'proper' prescribed formula for their HSUI argued that developing a HSUI 'properly' involved far less discretion than alternatives and ultimately produced a more valid and reliable instrument. Unsurprisingly, the 'proper process' group and the 'opportunistic' group had very different views of whether 'opportunistic' instrument construction was acceptable. HUI developers, who described devising their (opportunistic) classification system as a *'cognitive exercise … filling out a space, a hypothesis space around what we thought the best way would be to represent health'* (HUIP2), were criticised for *'forcing the data to fit a strict model'* which some perceived was not valid (SFP1). With respect to (also opportunistic) EQ-5D, one developer suggested:

> (my impression) … was very strongly, that those guys sat in a room and just decided which they thought were the best items to include and did no analysis whatsoever to look at whether they had the right type of coverage of items … (AQoLP2).

In contradiction to this, an EQ-5D developer asserted:

> It wasn't completely just a bunch of people sitting down in a room and coming up with their own ideas about what the domains ought to be. I know that there's sort of an impression that that was the case but there was very early research done passing around of that lay concept of health (EQP5).

Funding sources and resource availability enabled or constrained the ability of some projects to achieve their goals, for example, by limiting (HUI) or enabling (AQoL) formal valuation studies. Some developers worked under budgetary or time constraints in producing their instrument and had limited capacity to improvise beyond the task. For HUIP2, for example, the core process can be characterised as *Steamrolling ahead, within constraints* (Table 4). This process was focused more on development than evaluation; team members described trying to *'be careful as we could from a developmental point of view, in terms of putting the thing together'* because they knew they did not have the resources to do any empirical testing (HUIP2). On the other hand, EQP5 joined the EQ-5D later and described their focus on *Staying ahead of the game,* and *'continuously refreshing what it's doing'* which was enabled by the instrument generating a good stream of revenue/business model: *'so that you've continuously got quite a large investment in R&D into the instrument and the method and so on'* (EQP5).

## 6. Diverse ideas about what makes a good HSUI

Each developer valued particular characteristics of their own HSUI and had different interpretations of what makes a 'good' instrument (and, in turn, a poor instrument) and how quality data should be obtained (Table 5 supplementary).

Understandably, each developer rated the performance of all instruments against their own personal conception of a 'good' instrument. These conceptions sometimes changed throughout the development process and over time. For example, psychometrics might become more important later in the development processes, or common sense decisions might be used to refine health states, following formal psychometric testing, *'because … it's not a purely statistical question and you have to have a common sense approach'* (SFP2). It also meant that we observed subtle variation in each developer's conceptualisation of validity or what constituted a valid instrument. For instance, in some cases, validity was conceptualised as the extent to which an instrument could reflect the lived experience of particular conditions of health and illness. For others, it was reflected in an instrument's ability to represent all possible health states or was contingent on making sense to clinicians.

### 6.1. Comprehensiveness and sensitivity, or simplicity and pragmatism

One important contrast in valued qualities was that between comprehensiveness and sensitivity on one hand, or simplicity and pragmatism on the other. As with the previous contrast we made between types of goals, this contrast in valued qualities was not absolute: there was a small amount of overlap between these valued qualities for some

instruments, but overall the contrast was strong and helped explain the differences between instruments. Prioritising qualities was in some cases a finely balanced *'trade-off'* (QWBP1) for a number of developers, as reflected in several instances of contrasting valued qualities amongst developers of the same instrument (Column 1, Table 5).

Comprehensiveness and sensitivity were highly prioritised qualities for several developers. A comprehensive instrument was characteristically longer to ensure full coverage of, and sensitivity to, a large number of health states. For developers who valued comprehensiveness, 'good' instruments captured *the full gamut of quality of life* (AQoLP2) and enabled comprehensive assessment of health state (versus minimal core coverage). Comprehensive HSUIs were justified as making it possible to reflect real world complexity, for example, in measurement of functions and symptoms. Such HSUIs were generally designed to fulfil population-level bureaucratic aims, thus needing to be sensitive to nuance in every possible health state to capture population-level variation, such that '*every person in the universe can find their own health state if wishing so'* (15DP1); or '*it's extremely rare that somebody has a problem that is not captured in one of those indexes'* (QWBP1).

In contrast, other HSUI developers valued simplicity and pragmatism: their process prioritised stripping back to basic health states, to locate *a common core* for easier valuation, '*so you either had problems or you didn't in a way. The minimum needed for the objective'* (EQP1). Some devised their classification systems with a focus on design features that ensured simplicity for the user/administrator of the instrument, for valuation. HUIP1 described their process of *'explicitly, deliberately limiting the number of attributes we included, to make the cognitive task of valuing the health space more manageable'* (HUIP1). This matters, one participant explained, because quality in the data is lost when people reach their information processing limits. EQP4's focus was *Improving user experience:* their aim was to provide a coherent instrument that was easy to use, '*both from the point of view of the person that's filling it in and the point of view of the user, whether they be pharma companies, national health services, or even individual clinicians'* (EQP4).

Developers mindful of simplicity and pragmatism recognised the value in existing work ('*I think we accepted that all these other instruments were out there and … they'd probably covered the dimensions that needed to be in there'* (EQP1); *'the amount of research that went into develop the SF-36 … was a really good starting point'* (SFP2) rather than striving for originality. SF-6D developers described *'Adding value'* and *'Applying common sense'* (Table 4), reflecting this quality. They detailed a process of working to maximise return on existing effort and available resources and building on that incrementally. This is in comparison to 15D, for example, whose developers started new, reasoning that they could not be sure that existing generic measures would capture the issues of importance to their target population.

Those interviewees that valued simplicity and pragmatism were 'baffled' by instruments that set up millions of health states, reasoning that empirical work demonstrates that fewer than one hundred health states have been observed in practice. They argued that a large number of dimensions leads to problems of overlap and confusion for people valuing the states.

Variation in the developers' perceived indicators of a 'good' instrument was particularly evident in how they prioritised and 'traded-off' conflicting HSUI criteria in line with their most valued qualities (e.g. a comprehensive classification system that could be sensitive to the greatest number of health states with the pragmatism of a short, simple system with fewer dimensions). But *'the problem of course is that these criteria are at least to some extent, conflicting … so you have to somehow strike a balance between these criteria … a fair balance between these different parts of criteria'* (15DP1). Many participants acknowledged that favouring one criterion mostly unbalanced or affected others negatively, so they had to make decisions either individually or as a group.

## 7. Selecting valuation techniques

Commitment to simplicity and pragmatism carried over into people's approach to valuation (Table 5 Column 4). Some developers selected their valuation methods with a focus on the user of the instrument – they described prioritising the simplest technique, or the more understandable method for lay populations. For example, 15DP2, who emphasised the importance of *Designing for ordinary people,* said of the VAS thermometer: '*everyone can grasp that, everyone can understand what it means. It's a visual and it is easy to use, it's been applicable, it's easy to comprehend. Other ways of deriving the values, I think they are much more difficult for lay people'* (15DP2). Other developers described their strategies for making their techniques more approachable; for example, HUIP1 described deliberately building in space for people to think before they choose.

> Doing the VAS first and then the Standard Gamble is a cognitive process of giving the person some time to think about it and reflect and then give you their answer on the choice based technique (HUIP1).

Most development teams made valuation technique decisions early in the process; sometimes these had been made earlier and became *'given'* (SFP1) (Table 3). Some conducted or turned to published empirical work comparing different scaling methodologies (e.g. EQ-5D developers tried person trade-off (PTO) but people didn't understand it). Other principles employed included following disciplinary tradition or doing what was 'fashionable' or publishable at the time: for example, the DCE method was described as a '*modern fad'* (EQP4). Others developed new methods because existing options could not do what they needed them to do (e.g. 15D, where there were so many health states to be valued that existing methods were insufficient). Even within instruments there were differing recounts and differences in opinion across developers; the EQ-5D, a HSUI developed by a relatively large group of people in different locations, is a good example. One developer described a process involving *a lot of discussion, about which one (technique) was best* and group consensus on an official valuation method, *because now and again, you just have to make decisions like that … Sometimes we just have to go with the main consensus even if we don't agree with it … that's the general ethos* (EQP1), while another had a contrasting view and referred to a 'kerfuffle' over valuation:

*They've (group members) kind of gone off and made their own scientific decision about it without having worked out a consensus in the EuroQol Group. I mean it was a bit of a kerfuffle … it was actually a bit controversial at the time* (EQP5).

There was disagreement over the importance of weighting techniques. Not all developers were happy with their valuation systems; some planned to change these, others did not have the resources to do so.

## 8. All HSUIs are the product of a host of variables, shaped by their environment, and thus perform uniquely in different contexts

It is not surprising, given the subtle and consequential decisions that were made (individually and collectively) throughout each instrument's development, that they each have their own strengths and capabilities. There was wide acknowledgement by the developers that the HSUIs perform uniquely when applied in different contexts, each with their own strengths and weaknesses. One concluded,

> There's no perfect instrument … in many ways, quality of life assessment is in a state of alchemy … and the instruments, I think … perform remarkably well given how different they are and how similar a result they come up with (QWBP2).

One HSUI may work well in one context (e.g. academic research) but not another (e.g. clinical practice). Most developers consequently emphasised that instruments should be selected to best fit with the goals of the user or to suit specific needs. One HSUI may be more

important for public health care decision making bodies, another for clinical studies, an academic research project, or a pharmaceutical company comparing different things, depending on the outcome of interest: *'it is a social question, not a technical or economic question, what it is we want to measure'* (AQoLP1). However, using different HSUIs in different situations rather than across all conditions for consistency will increase validity, but with implications for comparability, '*It turns out that maybe things, they're not comparable'* (QWBP1),

> If you start using different instruments in different situations, you know, you'll run into questions about well can you really compare QALYs generated by different instruments and the answer probably is, is no but if you don't think one instrument is adequate in all situations, you're kind of forced to do that in some situations (SFP2).

> I do think that the validity of instruments to some extent depends on condition and that contradicts the fundamental aim of a lot of economists is that they want one instrument across all groups, so you have this sort of … it makes a conflict or problem (SFP2).

## 9. Future instrument development

The developers were invited to provide advice for future researchers contemplating further development of HSUIs. Here we summarise their key points.

Most developers said that progress in the field was contingent on funding. For example, EQ-5D began as an unfunded research group with no formal organisation but is now in a position to charge a license fee for commercial use (while being free to academics and the public), so has capacity for '*continuously refreshing what it's doing and staying ahead of the game*' (EQP5). HUI developers indicated that if funding were available, they would be interested in examining Item Response Theory (IRT) and Computer Adaptive Testing (CAT) in relation to instrument development.

Some developers emphasised the need to continue to develop methods and expand disciplinary reference points. One, for example, said: '*I think if anybody's literally going to go out there and do this again, then they would have a much better econometric toolbox'* (SFP1). Others suggested moving away from econometric traditions and towards psychology in particular, some suggesting economists have been '*naïve in the way they thought people process information'* (AQoLP2). This included, for example, paying attention to criteria in descriptive systems rather than criteria for utility weights, '*looking at what psychology has to offer'* to measure traits (AQoLP1) and focusing on psychometrics, specifically examining the sensitivity of HSUIs to happiness to provide essential insights to the existing HSUIs:

> What should happen is we ought to be looking at what psychologists have had to offer and asking ourselves the question, "If utility doesn't correlate with happiness, why?" Is it that we don't want happiness or is it that there's something wrong with our instrument? (AQoLP1)

Relatedly, there was some advocacy for increasing the use of qualitative research techniques to understand the impact of diagnoses on people's lives, incorporating social care into the scope of instruments, developing a children's mental health instrument, encompassing health and non-health sectors, or to track service use and compare across conditions and treatment settings.

Perhaps most controversially, one developer - while maintaining that their team probably made the right choices at the time - had come to believe that valuation methods used in HSUIs are irreparably flawed. He therefore rejected the use of HSUI, the calculation of QALYs, and making policy decisions based on imagined health states, arguing that: '*I think it's really a serious problem of misallocation, if we base things on people's misguided representations of what the future would be like'* (EQP5).

## 10. Discussion

Descriptive and empirical studies demonstrate that 'generic' HSUIs perform differently in head-to-head comparisons and this is widely acknowledged in the published literature (e.g. (Richardson et al., 2014; Brazier et al., 2017a,b)). But, in practice, once they are used in the calculation of QALYs those differences seem to be forgotten or become less apparent, and the resulting values (0–1) are commonly applied as if they are homogenous and comparable. That is, one QALY generated using one HSUI appears to be very much like one QALY generated using a different HSUI. Ultimately, their generic label conceals the deep processual, conceptual, and disciplinary differences sitting behind each HSUI.

Using unique data collected from developers of all the major HSUIs, we highlight the meaning and social value underlying HSUIs that is absent in existing quantitative evaluations of these measures. This analysis captured the challenge and practical realities of developing a generic instrument, and illuminated the central role of human, normative judgments in addition to technical considerations. The developers described a chain of decisions that they or their team made throughout the development process, guided by distinct explicit or implicit purposes and personal views of what the HSUI should do or achieve. Teams of developers made different sets of decisions, such that the HSUI that underpins each QALY calculation is very different. Our qualitative findings highlight that the construction of a HSUI does not follow a preconceived formula: they are developed through complex human processes driven by differing values. The development of each HSUI was organised around unique purposes - for example, to measure children's cancer outcomes (HUI), the health of nations (QWB), or the effectiveness of hospital interventions (15D) – so they essentially measure different constructs. We also found differences in approaches used to define and develop the descriptive system, preference weighting methods, and diversity in developer perceptions of what qualities make a good measure, and the depths of those differences were marked.

Participants were key informants with detailed knowledge who provided in-depth data. We took care to recruit participants from initial and subsequent stages of development. It is important to note that because most instruments were developed 20–40 years ago, some participants were concerned about their ability to remember accurately and provide a complete account of all aspects of instrument development. However, all interviews were extensive and it was striking how powerfully developers remembered the goals and values that underpinned their work. Given a number of the key developers are retiring, this study is particularly timely. Our analysis was based only on retrospective interviews; although beyond scope for this project, we note that archival historical research, if documents from development processes were available, would complement our analysis. We consider it a strength of our sampling strategy that we spoke to a large proportion of living developers, including in some instances all developers of an instrument.

In this study, we have delved deeper than psychometrics or statistical properties captured in head-to-head comparisons (which can show, for example, that instrument choice may have a significant effect on health state estimates and effect size calculations). Distinctively, we spoke directly with developers of six HSUIs and have helped to explain *how* and *why* these different ways of producing these HRQoL or utility weights came to be different.

These findings are important because differences in health state values derived from different preference-based measurement instruments have been shown to have potentially important implications for QALY calculations, and in turn implications for public policy and resource allocation: the incremental cost per QALY of health care interventions and hence the cost-effectiveness of interventions. For example, a health technology can appear more or less cost-effective depending on the HSUI that is used. Our results help explain possible reasons for empirical differences in values generated across HSUIs and bring

further attention to possible differences in resulting QALYs.

Understanding each instrument's origins and how and why key decisions were made throughout their development is expected to be useful to a number of stakeholders including researchers contemplating using a specific HSUI, policy makers/funders, and future developers. Taking each in turn, our results may help prospective users to choose which HSUI to use in different contexts. For example, by alignment between prospective users' research goals and instrument goals, or based on greater understanding of the sensitivities or perceived strengths of particular instruments.

Results can also help policy makers and funders to better understand the data that result from the use of HSUIs and provide further complementary evidence regarding why a QALY is not a QALY. Finally, this study provides valuable lessons and insights for researchers contemplating development of a new instruments for use in health economics research drawn from interviews with developers with over 40 years' experience in the field. New instruments to which these insights are likely to be relevant include new disease and age specific HSUIs as well as those moving beyond HRQoL and QALYs to generate more encompassing measures of quality of life or wellbeing. This is in recognition of outcomes that matter to individuals and decision makers beyond health (e.g. socioeconomic status, home circumstances, social care) that HSUIs do not adequately capture. This is important both within the health sector and of course when considering resource allocation across sectors. This highlights the relevance of this study to inform not only current HSUIs but potentially also future quality of life and wellbeing instrument development.

## 11. Conclusion

The often stated assumption that “a QALY is a QALY is a QALY” has repeatedly been challenged in the literature, most often because of evidence suggesting the value attached to QALYs can differ based on the beneficiaries of the QALYs (e.g. 1 QALY generated for treatment of children might be valued differently by society or HTA committees to 1 QALY generated in treatment for the elderly) – e.g. (Lancsar et al., 2011; Gu et al., 2015) - but also due to recognition that different HSUIs produce different HRQoL (or health state) utility values applied in the calculation of QALYs. In this paper we provide further complementary evidence against the assumption, by offering new insight into how and why the differences in HSUIs have come about, and this a new understanding of why a QALY is not a QALY is not a QALY.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.socscimed.2019.112560.

## References


Bleichrodt, H., Quiggin, J., 2013. Capabilities as menus: a non-welfarist basis for QALY evaluation. J. Health Econ. 32 (1), 128–137.

Brazier, J., Ara, R., Rowen, D., Chevrou-Severac, H., 2017a. A review of generic preference-based measures for use in cost-effectiveness models. PharmacoEconomics 35 (1), 21–31.

Brazier, J., Ratcliffe, J., Saloman, J., Tsuchiya, A., 2017b. Measuring and Valuing Health Benefits for Economic Evaluation. Oxford university press.

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

Brazier, J., Tsuchiya, A., 2015. Improving cross-sector comparisons: going beyond the health-related QALY. Appl. Health Econ. Health Policy 13 (6), 557–565.

Brooks, R., Group, E., 1996. EuroQol: the current state of play. Health Policy 37 (1), 53–72.

Charmaz, K., 2006. Constructing Grounded Theory: A Practical Guide through Qualitative Analysis. Sage.

Coast, J., Flynn, T.N., Natarajan, L., Sproston, K., Lewis, J., Louviere, J.J., Peters, T.J., 2008. Valuing the ICECAP capability index for older people. Soc. Sci. Med. 67 (5), 874–882.

Coast, J., Kinghorn, P., Mitchell, P., 2015. The development of capability measures in health economics: opportunities, challenges and progress. Patient-Patient-Cent. Outcomes Res. 8 (2), 119–126.

Devlin, N.J., Brooks, R., 2017. EQ-5D and the EuroQol Group: past, present and future. Appl. Health Econ. Health Policy 15 (2), 127–137.

Drummond, M.F., Sculpher, M.J., Torrance, G.W., O'Brien, B.J., Stoddart, G.L., 2005. Methods for the Economic Evaluation of Health Care Programmes. Oxford university press, Oxford.

Ebrahim, S., 1995. Clinical and public health perspectives and applications of health-related quality of life measurement. Soc. Sci. Med. 41 (10), 1383–1394.

Finch, A.P., Brazier, J.E., Mukuria, C., 2018. What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. Eur. J. Health Econ. 19 (4), 557–570.

Fryback, D.G., Palta, M., Cherepanov, D., Bolt, D., Kim, J.-S., 2010. Comparison of 5 health-related quality-of-life indexes using item response theory analysis. Med. Decis. Mak. 30 (1), 5–15.

Given, L.M., 2008. The Sage Encyclopedia of Qualitative Research Methods. Sage Publications.

Green, J., Britten, N., 1998. Qualitative research and evidence based medicine. Bmj 316 (7139), 1230–1232.

Gu, Y., Lancsar, E., Ghijben, P., Butler, J.R., Donaldson, C., 2015. Attributes and weights in health care priority setting: a systematic review of what counts and to what extent. Soc. Sci. Med. 146, 41–52.

Hawthorne, G., Richardson, J., Day, N.A., 2001. A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. Ann. Med. 33 (5), 358–370.

Hawthorne, G., Richardson, J., Osborne, R., McNeil, H., 1997. The Australian Quality of Life (AQoL) Instrument: Initial Validation. Centre for Health Program Evaluation Melbourne, Vic.

Herdman, M., Gudex, C., Lloyd, A., Janssen, M., 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). Qual. Life Res. 20 (10), 1727–1736.

Kaplan, R.M., Anderson, J.P., 1988. A general health policy model: update and applications. Health Serv. Res. 23 (2), 203.

Kaplan, R.M., Sieber, W.J., Ganiats, T.G., 1997. The quality of well-being scale: comparison of the interviewer-administered version with a self-administered questionnaire. Psychol. Health 12 (6), 783–791.

Karimi, M., Brazier, J., 2016. Health, health-related quality of life, and quality of life: what is the difference? PharmacoEconomics 34 (7), 645–649.

Keeley, T., Al-Janabi, H., Lorgelly, P., Coast, J., 2013. A qualitative assessment of the content validity of the ICECAP-A and EQ-5D-5L and their appropriateness for use in health research. PLoS One 8 (12), e85287.

Killewo, J., Heggenhougen, K., Quah, S.R., 2010. Epidemiology and Demography in Public Health. Academic Press.

Lancsar, E., Wildman, J., Donaldson, C., Ryan, M., Baker, R., 2011. Deriving distributional weights for QALYs through discrete choice experiments. J. Health Econ. 30 (2), 466–478.

Longworth, L., Longson, C., 2008. NICE Methodology for Technology Appraisals. Springer.

Lorgelly, P., Lorimer, K., Fenwick, E., Briggs, A., 2008. The Capability Approach: Developing and Instrument for Evaluating Public Health Interventions. Section of Public Health and Health Policy. University of Glasgow.

Netten, A., Burge, P., Malley, J., Potoglou, D., Towers, A.-M., Brazier, J., Flynn, T., Forder, J., 2012. Outcomes of social care for adults: developing a preference-weighted measure. Health Technol. Assess. 16 (16), 1–166.

Patton, M., 2002. Qualitative Research & Evaluation Methods, 3ed. pp. 91320 Thousand Oaks, California.

Richardson, J., Iezzi, A., Khan, M.A., Chen, G., Maxwell, A., 2016. Measuring the sensitivity and construct validity of 6 utility instruments in 7 disease areas. Med. Decis. Mak. 36 (2), 147–159.

Richardson, J., McKie, J., Bariola, E., 2014. Multiattribute Utility Instruments and their Use.

Richardson, J.R., Peacock, S.J., Hawthorne, G., Iezzi, A., Elsworth, G., Day, N.A., 2012. Construction of the descriptive system for the assessment of quality of life AQoL-6D utility instrument. Health Qual. Life Outcomes 10 (1), 38.

Simon, J., Anand, P., Gray, A., Rugkåsa, J., Yeeles, K., Burns, T., 2013. Operationalising the capability approach for outcome measurement in mental health research. Soc. Sci. Med. 98, 187–196.

Sintonen, H., Pekurinen, M., 1989. A generic 15 dimensional measure of health-related quality of life (15D). J. Soc. Med. 26 (1), 85–96.

Stevens, K., 2017. Using Qualitative Methods to Develop a Preference-based Health-Related Quality of Life Measure for Use in Economic Evaluation: the Development of the Child Health Utility 9D. Qualitative methods for health economics. J. Coast. London. Rowman & Littlefield International Ltd.

Tolley, K., 2009. What Are Health Utilities. Hayward Medical Communications, London.

Torrance, G.W., 1987. Utility approach to measuring health-related quality of life. J. Chronic Dis. 40 (6), 593–600.

Torrance, G.W., Boyle, M.H., Horwood, S.P., 1982. Application of multi-attribute utility theory to measure social preferences for health states. Oper. Res. 30 (6), 1043–1069.

Torrance, G.W., Feeny, D.H., Furlong, W.J., Barr, R.D., Zhang, Y., Wang, Q., 1996. Multiattribute utility function for a comprehensive health status classification system: health utilities index mark 2. Med. Care 34 (7), 702–722.

Weinstein, M.C., Russell, L.B., Gold, M.R., Siegel, J.E., 1996. Cost-effectiveness in Health and Medicine. Oxford university press.

Wilson, P., 2010. A Comparison of Health State Utility Measures across Criteria, Informed by Previous Literature, Economic Theory and Interviews with Key Developers of the Instruments. Honours Thesis. BA (Economics). Newcastle Univesity Business School.