A multi layered "Deep Dive" measure of health-related quality of life based on the EQ-5D: Introducing the approach

Brendan Mulhern[1], Aureliano Finch[2], Bas Janssen[2], Mark Oppe[3]

1 Centre for Health Economics Research and Evaluation, University of Technology Sydney

2 EuroQol Research Foundation, Rotterdam

3 Maths in Health, Rotterdam

Abstract

The EQ-5D descriptive system includes dimensions that are measured using a single item as concise health state descriptions are required for valuation. Such an approach facilitates the generation of utility values but limits the possibility of collecting rich profile information about the patients' status on each dimension. In some clinical settings a more detailed understanding of the impact of a condition on a particular health dimension is required, which in turn may limit the use of EQ-5D for those conditions. To address this gap the DSWG has embarked on a pilot study in which the potential for developing an instrument that combines the benefits of both preference-accompanied and profile measures is investigated. This focuses on the development of Dimension Specific Modules for the EQ-5D, and is therefore referred to as a multi layered 'Deep Dive' approach to measuring health and quality of life.

The Deep Dive approach includes a higher-level preference- accompanied measure (i.e., the core EQ-5D descriptive system), and a set of items associated with each dimension (the Dimension Specific Modules). The Deep Dive program of work would involve the identification and development of items measuring each dimension to be included in the Dimension Specific Modules.

The aim of this conceptual paper is to introduce the Deep Dive approach and outline the methods that could be used to develop such an instrument. This is done across four sections. Section 1 introduces the basis of and justification for the Deep Dive approach. Section 2 outlines the methods that could be used to develop Dimension Specific Modules, and a range of methodological issues that would require understanding and investigation. Section 3 discusses the scientific and strategic advantages and disadvantages of the approach, and Section 4 provides an overall discussion including questions for the EuroQol membership, and outline the potential next steps for this programme of work

Funding: The Deep Dive pilot study is funded by the EuroQol Research Foundation. The views of the authors are not necessarily those of the funder. The authors are all current or past Descriptive System Working Group (DSWG) members. They thank the rest of the DSWG for their input.

Introduction

The EQ-5D is the most widely used generic multi attribute utility instrument (MAUI) worldwide (Wislof 2014; Rowen et al 2017; Kennedy-Martin et al 2020). EQ-5D was originally developed to accompany condition-specific instruments, and provide values for the calculation of quality-adjusted life years (QALYs) to inform resource allocation decision making (Brooks, 2013). Since the instrument was developed to be used as a preference-accompanied measure, the descriptive system was intentionally kept brief, with the EQ-5D-3L including five dimensions and three levels, therefore describing 243 health states. The EQ-5D-3L has been widely used in informing cost effectiveness analyses, and as part of clinical trials, observational studies, population health surveys, and as a Patient-Reported Outcome Measure (PROM) as a part of routine outcome measurement (Devlin & Brooks 2017; Devlin & Appleby 2010).

The challenge for developers of generic MAUI descriptive systems is to describe health in generic terms, using a limited number of dimensions and items within dimensions that display a high level of validity and sensitivity to change across diverse health conditions, without resulting in excess respondent burden or in a complex descriptive system not amenable for valuation. However, since the number of unique health states that individuals may experience it is infinitely high, it is inevitable that no generic MAUI will be valid and responsive across all conditions, either because an important dimension is omitted or changes in health are not identified. For example, the EQ-5D-3L has demonstrated good measurement properties across a wide number of conditions and populations (Finch et al 2018; Longworth et al 2014). However, the concise descriptive system results in content validity limitations, and there is evidence to suggest that not all relevant aspects of health and disability in certain conditions are captured (Shah et al 2017; Brazier et al 2014). Furthermore, several studies have found that the EQ-5D-3L exhibits a ceiling effect when compared to more granular instruments that include more items and response levels, and broader domain coverage (Brazier et al 2004; Bharmal and Thomas 2006).

In response to these issues the EuroQol Group has investigated ways to increase the sensitivity of the instrument. This was done first by increasing the number of response options (levels) from 3 to 5, which resulted in the development of the EQ-5D-5L (Herdman et al, 2011). A recent review of the performance of the EQ-5D-5L across 99 papers established that the instrument has strong psychometric properties across a broad range of populations, conditions and settings (Feng et al 2021).

A way of improving content validity is to add, or 'bolt-on' one or more dimensions to the core descriptive system, and this has been done in numerous health areas. The descriptive impact as well as the effect of bolt on dimensions on values have both been tested, with varying results (see recent review by Geraerds et al 2021). The bolt-on research agenda has many conceptual and practical issues, including whether the additional dimensions can be generic or condition specific, how many dimensions can be added to the core descriptive system, and how adding further dimensions impacts the valuation process, and the values elicited. In addition to this, there is a conceptual issue around what constitutes a bolt-on, as certain bolt-ons could be defined as nested underneath an existing EQ-5D dimension, or be a health consequence of an existing dimension. For example, itching (from the psoriasis bolt-on set, Swinburn et al 2013) could be considered as a type of discomfort, and therefore might better be nested as a subset of discomfort rather than added alongside the core descriptive system. This could avoid double counting, and complex interactions in both self-report measurement and valuation. A similar issue relates to the composite dimensions included in the EQ-5D (PD, AD), and evidence investigating what aspects of these patents are responding to is inconclusive.

Is there an alternative approach?

Another way to increase the sensitivity and information provided by the EQ-5D lies in conceptualising the EQ-5D as a multi-level hierarchical instrument (See Fig. 1). The utility values reside at the top level (Utility level), summarizing the impact of the core five dimensions, that form the middle level (Dimension level). This can be complemented by adding a new "level" to the instrument. This level includes a set of dimension-specific items (or Dimension Specific Modules (DSM)) nested below each of the five core dimensions (hence the definition as the Module Level). This approach retains the EQ-5D descriptive system as the core while enabling a 'deep dive' into each of the five dimensions using the DSMs. It is a concept that could underpin future evolvements of the EQ-5D instruments.

The aim of this conceptual paper is to provide a broad overview of the Deep Dive approach to the EuroQol membership. This is done across the four sections outlined below.

- Section 1 will introduce the Deep Dive approach
- Section 2 will outline the methods that could be used to develop DSMs, and a range of methodological issues that would require understanding and investigation
- Section 3 will discuss the scientific and strategic advantages and disadvantages of the approach
- Section 4 will provide an overall discussion including questions for the EuroQol membership, and outline the potential next steps for this programme of work

The Descriptive Systems Working Group (DSWG) is conducting pilot work into this novel approach, and this paper serves as a broad summary of some of that work. The aim of this paper is to stimulate discussion and feedback from the membership. The authors and the DSWG also hope that the paper acts to stimulate interest into the idea, and the many conceptual and methodological challenges, and results in further collaborations to investigate the approach and the ideas raised in this paper further.

Section 1 – Introduction to the Deep Dive approach

The structure of the Deep Dive and terminology used

As mentioned above, the Deep Dive approach conceptualises the EQ-5D as a multi-level hierarchical instrument, with three distinct but interrelated levels (See Fig 1). Each level is defined as:

- Utility level – Existing EQ-5D value set generated from the core descriptive system (or dimension level)
- Dimension level – The core five dimensions of the EQ-5D descriptive system
- Module level: DSM including sets of items that measure the construct(s) assessed by each core dimension in more depth and detail.

Previous research and existing instruments that inform the approach

Previous research has conceptualised instruments as multi-levelled, and the hierarchical instrument structure has been implemented in the development of an earlier instrument. The Quality of Care in Dialysis Centres Questionnaire (QCDQ, Oppe et al 2005, see Figure 2) used this approach, and included an overall visual analogue scale as Level 1. This was followed by a set of four satisfaction items (level 2) asking about level of satisfaction with different aspects of care and practitioners, including doctors, nurses, other staff members, and the facilities. Level 3 was a set of eight descriptive items below each satisfaction item.

There are other measures of quality of life, and MAUIs, that adopted a hierarchical structure. For example, the Assessment of Quality of Life (AQOL-8D; Richardson et al 2014) system conceptualises the structure as two physical and psychological 'super dimensions' that consist of eight dimensions (covering independent living, pain, senses, mental health, happiness, coping, relationships and self-worth). Each dimension in turn includes a set of items (35 in total). Figure 3 displays the full structure which is in line with the conceptualisation of the EQ-5D as a multi-level hierarchical instrument, the difference being that respondents to the AQOL-8D complete the full 35 items.

The SF-36 system is another informative example. The 36 items are nested within eight dimensions (physical functioning, role physical, role emotional, social functioning, pain, mental health, vitality, general health) which can be transformed into two summary scores (Mental Component Summary and Physical Component Summary), and which are not framed as individual items but provide a broader level of information. In the opposite direction, a subset of the 36 items have been condensed into a six-dimension MAUI (SF-6Dv1 and SF-6Dv2; Brazier et al 2002; Brazier et al 2020; Mulhern et al 2020). This was done by selecting items to represent the dimensions using psychometric approaches to ensure coverage across the severity range of the dimension. The difference between the SF-36 system and the deep dive approach is that the SF-36 profile measure was used as the basis for the development of the MAUI descriptive system (rather than the MAUI being used as a basis for the development of an in depth set of items about the construct measured by the dimension

Although some MAUIs can provide guidance, this is a new area of exploration for the EuroQol Group as the multi-level approach not been previously considered as a methodology to extend a generic MAUI with an additional profile measure.

Example – Mobility/physical functioning

As an example of a single core dimension as part of the multi-level approach, consider Mobility (MO). The utility level consists of the utility decrements assigned to the responses to the dimension (i.e., the dimension level) which is framed as the amount of problems with 'walking about' across three or five severity levels. The development of the DSM at the module level would include items assessing other areas of mobility and physical functioning. By developing a mobility DSM, we will be able to detect whether the patient has problems with walking different distances, or with other aspects of mobility such as bending or kneeling, climbing stairs, or lifting and carrying, and how severe each of these problems are. The DSM would be developed using in depth and established instrument development methods to ensure that the items have content and construct validity, and can be scored appropriately (see Part 2).

Section 2 – Methodological approaches to develop, score and value a DSM

General introduction

There are a number of key areas where different methodological approaches could be used to develop DSMs. These are the generation of items for inclusion in the DSM, linking the DSM to the 'core' dimension and scoring the core dimension and DSM. Existing methods used for the development, assessment and scoring of Patient Reported Outcome Measures (PROMs) (see Brazier et al (2022); Mokkink et al (2010)) can be adapted for the generation of DSMs, with patients and populations where the DSM will be applied being central to the development. The development would be in line with EuroQol Group guidance around the IP status of instruments.

Each DSM should include sufficient items to fully measure the dimension for which it is being developed, as would be the case in developing a PROM to assess each dimension (taking into account the need to minimise respondent burden). However, the number of items for each DSM does not need to be the same. Other considerations in the development of DSMs are highlighted in the 'Methodological and conceptual issues to consider' subsection below.

Methodological approach to developing a DSM

Developing items

Prior to embarking on the development of items, theoretical work to fully understand differences in the health dimensions of the EQ-5D, and what is being measured by each, would be required. This would provide the basis for understanding how broad or narrow the focus of the EQ-5D dimensions is, and inform the constructs within the dimension that could be measured.

Items for the DSM could be generated de novo but could also be based on items included in existing instruments (in line with the development of other instruments, for example Brazier et al (2022); Carlton et al (2022)). Item generation would be based on qualitative work and literature reviewing and identification and adaptation of relevant items from existing instruments. Following the development of an item pool, important decisions such as format, response levels and wording would be considered at this stage, and tested throughout. A key issue is whether items need to be adapted to fit the structure of the core EQ-5D dimension, as is recommended for bolt-on development (Mulhern et al 2022). Another key issue would be consideration of how the items are related to the dimension of interest, and this would also be tested. A further qualitative phase to refine and reduce the item pool taking these issues into account would also be conducted. This would be adaptable for

the DSM, but could involve patient focus groups, expert input and semi structured face validity interviews. The result of this phase would be an item pool for further testing and refinement.

Taking the example of mobility/physical functioning, items could be identified from existing instruments dimensions such as the SF-36 Physical Functioning dimension, or relevant PROMIS items, or developed from initial qualitative work. Further qualitative work in populations with a range of mobility problems would be conducted to examine the face and content validity of the item pool. Any redundant items, or items displaying limited evidence of face validity, would be removed at this stage.

Testing and selecting items

The development of the shortened item pool would be followed by a psychometric phase to test and select items for the DSM. The refined item pool would then be administered alongside other instruments, including the EQ-5D-5L, to relevant patient and population samples. This could be done online via patient panels, or in clinical settings. Psychometric analysis, with the aim of developing the 'final' item set for the DSM would be conducted. This would be done by using both classical test theory (CTT) psychometric methods and modern test theory (IRT). Capellieri et al (2014) review the use of CTT and IRT for evaluating outcome measures, and conclude that both CTT and IRT are valid to support the maximisation of the content validity of different instruments.

CTT would be used to examine item performance using established criteria (see e.g., Fayers and Machin, 2016). This would include item acceptability indicators such as missing data, response level use, and floor and ceiling effects. Construct validity, including convergent and known group validity would also be assessed. A range of dimensionality assessment techniques including factor analysis, principal component analysis and structural equation modelling would be used to assess the loadings of the items to the different EQ-5D dimensions, and internal consistency reliability (using Cronbach's alpha) could be examined based on the DSM items identified. This would allow tests of assumptions related to the extent to which items are equally related to a unidimensional or multidimensional latent trait.

Modern Test Theory methods (focused on IRT approaches) would also be used to examine item performance. IRT methods (summarised by Fayers and Machin (2016) and Edelen & Reeve (2007)) are a set of theoretical approaches and associated practical methods used for the construction of measurement instruments [76]. IRT models link observed item responses to respondents' location on an unmeasured underlying latent trait (described as the 'theta' (θ) scale). The theta scale builds on

the assumption that a set of items measure a unique and identifiable continuous latent trait, and models an unobservable continuous dimension that is assumed to be unidimensional (i.e., measuring one construct). Applying an IRT to a set of items is said to 'calibrate' those items on the unidimensional theta scale. In the case of patient-reported outcomes theta represents a dimension of health of QoL that is measured by responses to a set of items assessing the same unidimensional concept. The scale represents a continuous severity range across theta. IRT is used to inform item performance by estimating item threshold parameters that represent the transition between levels, and a slope parameter that provides a single figure estimate of how particular items discriminate at different levels of theta (and is a function of the threshold parameters and theta). IRT also allows for an assessment of item redundancy, differential item functioning, and fit to the dimension. IRT will be used to assess these indicators for each item (and the overall dimension coverage) for each DSM. IRT can also be used to score dimensions (see below).

Taking the mobility/physical functioning dimension, CTT and IRT methods could be applied to primary data collected on the DSM item pool and used to further refine the items included to produce the final DSM. Section 4 provides a practical example of this.

Linking and scoring the DSM

Following the development of a psychometrically valid DSM, the items included in the DSM need to subsequently be linked to the appropriate EQ-5D dimension. A single DSM, or DSMs across all five core EQ-5D dimensions can be used as stand-alone generic PROMs (or sets of PROMs) while not affecting responses to the core EQ-5D, or therefore the estimation of values. The DSMs also require scoring for use in the assessment and comparison of patient and population health, and multiple methods could be used to test and develop a number of scoring systems with different purposes. Both linking and scoring methods would build on existing approaches used for the development of PROMs. The link between the DSM and core EQ-5D-5L dimension and utility values for that dimension can be explored using psychometric methods (for example exploratory and confirmatory factor analysis) and in principle be established via mapping (response or utility mapping). IRT could also be used to establish a direct link between the DSMs and their respective EQ-5D dimensions. This could also facilitate the development of computerized adaptive testing (CAT) approaches to administering the DSM.

Scoring will also be developed based on approaches used for other PROMs. Scoring approaches for the items included only in the DSM, and also for the DSM plus the associated core EQ-5D-5L dimension

will be explored. This includes simple approaches such as summary scoring, and total scores transformed onto comparable scales if DSMs include different numbers of items (for example SF-36 dimension scores developed by Ware et al (1992)). Scoring using psychometric techniques including Mokken Scaling could also be explored as part of the DSM linking analysis. This approach has been tested for both the EQ-5D-5L (Feng et al 2022) and EQ-HWB (Feng et al, 2021). IRT can be used to develop calibrated theta scores/summary scales for each DSM, incorporating the original dimension.

For both linking and scoring, patient data would be needed to understand how each overall DSM, and items included, are related descriptively and psychometrically to the respective EQ-5D dimension (or dimensions if there is overlap). Depending on the items selected for each DSM there may be a structural difference in how they are linked to the EQ-5D items, and the links between reported DSM problems and EQ-5D dimension scores need to be investigated thoroughly, to make sure that the correct links are used to connect the DSMs and the EQ-5D. Of course, these issues related to the links between the DSMs and the EQ-5D dimensions might be mitigated by taking this into account when selecting the items for the DSMs (for example by developing DSMs linked to single core dimensions, but with item subsets that cross load across core dimensions).

Methodological and conceptual considerations in developing a DSM

There are a range of broad methodological and interlinked conceptual issues to consider in the development of DSMs. A number of these will be discussed below.

One key conceptual issue that will need to be considered in the methodological development of items is what is measured by the core EQ-5D dimensions, and implications that would have for items included in DSMs. As an example, it has been found that in some cases, MO and UA appear in the same underlying factors, and therefore items in these DSMs could also overlap in terms of what is being measured (Finch and Mulhern, under review). This raises the question of to what extent it is valid to have similar items in different DSMs, or whether DSMs can be developed so that certain items align with multiple DSMs. If the items have a content and psychometric link back to the original dimension, then it could be argued that conceptually these items can be included in multiple DSMs if the relationship exists.

An associated issue, and a core conceptual challenge for a DSM is whether the content should only be aligned with the current conceptualization of what is measured by the core dimensions, or whether

this should be broadened. This is specifically relevant for SC and AD, for example should a DSM include broader issues such as eating (for SC), or other mental health concerns, or impacts of those concerns, for AD. There would not be a direct link to the core EQ-5D anymore but content validity (a key requirement of valid instruments), and sensitivity to the key impacts of a dimension, would be increased. This also raises the issue of whether DSMs should be generic, and therefore improving measurement across conditions, or whether they can apply to specific conditions.

Another consideration in the development of a DSM is the role of existing EuroQol instruments with experimental and beta status, including the EQ-HWB (Brazier et al 2022). It is possible that existing EQ-HWB items could be part of certain DSMs given the overlap in conceptual coverage. This could also apply to bolt on dimensions where some overlap is possible. Work is ongoing to establish the relationship between EQ instruments, and this could be considered in light of the development of DSMs, with further work also conducted. Equally, DSMs could be developed as part of the EQ-HWB-S using the items included in the long form as part of the DSMs where relevant.

There are a range of considerations linked to the structure, format and wording used that would also be considered in the development of any DSM. Recent guidance has specified a set of criteria for the development of bolt on dimensions (Mulhern et al 2022), and similar guidance could be developed for DSMs. For example, should items in a DSM be specified to have the same item structure as a core dimension (as specified for bolt-ons (Mulhern et al 2022)), or could this be more flexible? This includes whether items should be specified to be worded as a set of statements, or allowing the use of Likert items where relevant. Another consideration, which impacts the measurement relationship between items, is whether using other response level categories than severity (for example frequency) is permissible. Similarly, another feature that influences the measurement relationship is the inclusion of both positively and negatively worded scales. The choice of recall period also interacts with these considerations (if the DSM broadens out to allow frequency response options, or uses EQ-HWB items where 7 days is used, is the core EQ-5D-5L recall of 'today' the most appropriate, or possible to use for all items?

Another item specific consideration is the inclusion of composite dimensions in the EQ-5D-5L (including pain and discomfort, anxiety and depression, and arguably self-care as it asks about both washing and dressing). The impact of composite descriptions on the measurement and valuation of those dimensions is already the subject of research (Tsuchiya et al 2019) but raises issues for the development of DSMs. For example, taking pain and discomfort, should separate dimensions be

developed to assess pain and discomfort? There are arguments for both approaches. A separate dimension for discomfort would allow for a range of discomfort specific issues that are not directly perceived as pain to be measured. Also, dimensionality assessment has suggested they are measuring different constructs (Engel at al 2020). However, there is likely a lot of overlap between what is perceived as pain and discomfort, and therefore a single dimension could be argued for, particularly if IRT type scoring was used to allow for flexible item pool administration based on CAT. Taking anxiety and depression, there are different indicators and severities of both anxiety and depression that could be assessed in individual DSMs, but also overlap in the constructs measured.

Section 3: Advantages and disadvantages of the Deep Dive approach

Scientific and strategic advantages

The development of DSMs has a number of potential benefits in the extending the measurement of health, and also for the usefulness of the data collected. PBMs by their nature are limited in the information that they can provide. This approach could broaden the applicability of the instruments across a variety of clinical decision making, research focused, and routine outcome measurement settings. For example, a single instrument could provide a utility value for use in resource allocation decision making, and also detailed information for each dimension covering a wide range of constructs for use in clinical settings to inform treatments. Strategically this potentially extends the market for the EQ-5D as both a profile and preference-based measure of health.

Scientifically, this approach also opens the potential for the development of a psychometrically valid instrument developed using state of the art psychometric methods and enables further assessment of the measurement characteristics of the core EQ-5D-5L dimensions, as well as the DSM. This is important, as not all users will apply the DSM, so it is essential to continue to build knowledge about the core EQ-5D dimensions.

Scientific and strategic disadvantages

There are also a number of scientific and strategic disadvantages that require careful consideration before embarking on the development of DSMs. The cost of developing scientifically rigorous DSMs will be relatively high, and although the initiative would extend the use of EQ-5D into new areas, the business case is currently unclear, and needs further consideration. The usage of each DSM also varies across populations and patient groups, and some DSMs may be more widely used and accepted than others (which has implications for the order in which DSMs should be developed).

There are many methodological choices for developing DSMs. To ensure that the development is not disadvantaged by methodological approaches, pilot work is required to examine which methods are most appropriate (and can be adapted for use across the dimensions).

Section 4 - Summary and Discussion

We have outlined a rationale for extending the EQ-5D by developing dimension specific DSMs, and also highlighted potential methodological approaches to do this. This approach would have the benefit of increasing the application of the EQ-5D in settings where profile PROM measures are often used. For example, the increase in information provided would be of use in clinical settings, and potentially clinical decision making (and alongside this, values based on the EQ-5D-5L can still be generated). The approaches used to develop DSMs would build on the latest qualitative and quantitative methods ensuring that the questionnaires were valid and reliable, and adding information to what is being measured at the dimension level. Linking the DSM to the dimension provides opportunities for a range of scoring approaches to be developed and used. However, users who only require values could choose to use the EQ-5D-5L descriptive system alone.

The limitations of the approach are a lack of clarity around the business case, and a series of methodological and conceptual challenges that require further understanding. Focusing on the core five dimensions still results in the limitation that some dimensions important in certain conditions are not measured. To counter this, the same approaches could be used for the development of bolt-ons and bolt on specific DSMs. There are also issues for valuation. In the first instance, the five dimensions are still included, so existing value sets can be applied. In future work, DSMs could be linked to values.

Our broader question for discussion by the membership is whether this initiative is seen as an area for further development by the group, and if so, how should we proceed strategically and scientifically?

Next steps for the Deep Dive initiative

The DSWG is currently conducting the pilot study which will include consultation with the EuroQol membership on both the rationale and justification for the approach, and the methods that could be used to develop a DSMs. Alongside this, extensive empirical analysis to test development approaches on existing data has been conducted, and will also form part of the consultation. If this initiative is then seen as of interest to the group and executive, a larger study will be proposed to consider the many methodological and conceptual questions highlighted above, and conduct further work to move

towards the development of a DSM considering the views of the membership, and recommendations on the appropriate way to proceed.

References

Akaike H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* 19: 716-23.

Bharmal M, Thomas J. (2006). Comparing the EQ-5D and the SF-6D Descriptive Systems to Assess Their Ceiling Effects in the US General Population. *Value in Health, 9(4):* 262-71.

Brazier J, Connell J, Papaioannou D, Mukuria C, Mulhern B, Peasgood T, Lloyd Jones M, Paisley S, O'Cathain A, Barkham M, Knapp M, Byford S, Gilbody S, Parry G. (2014). A systematic review, psychometric analysis and qualitative assessment of Generic Preference-Based Measures of Health in Mental Health Populations and the estimation of mapping functions from widely used specific measures. Health Technology Assessment, 18:34.

Brazier J, Peasgood T, Mukuria C, Marten O, Kreimeier S, Luo N, Mulhern B, Pickard AS, et al. (2022). The EQ Health and Wellbeing: Overview of the Development of a Measure of Health and Wellbeing and Key Results. *Value in Health*, https://doi.org/10.1016/j.jval.2022.01.009.

Brazier JE, Roberts J, Tsuchiya A, Busschbach J. (2004). A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Economics, 13(9):* 873-84.

Brooks R. (2013). *The EuroQol Group after 25 years*. Dordrecht: Springer.

Kennedy-Martin M, Slaap B, Herdman M, van Reenen M, Kennedy-Martin T, Greiner W, Busschbach J, Boye K. (2020). Which multi-attribute utility instruments are recommended for use in cost-utility analysis? A review of national health technology assessment (HTA) guidelines. *European Journal of Health Economics, 21:* 1245-57.

Cappelleri JC, Lundy J, Hays R. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics.* 2014; 36(5): 648-62.

Carlton J, Peasgood T, Mukuria C, Connell J, Brazier J…Mulhern B, Greiner W, Pickard AS, Augustovski F. (2022). Generation, Selection, and Face Validation of Items for a New Generic Measure of Quality of Life: The EQ Health and Wellbeing. *Value in Health*, https://doi.org/10.1016/j.jval.2021.12.007.

Devlin NJ, Appleby J. (2010). *Getting the Most out of PROMs: Putting Health Outcomes at the Heart of NHS Decision-Making*. London: King's Fund.

Devlin NJ, Brooks R. (2017). EQ-5D and the EuroQol Group: Past, Present and Future. *Applied Health Economics and Health Policy, 15(20):* 127-37.

Edelen MO, Reeve BB. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research,* 16(S1): 5-18.

Engel L, Haagsma J, Janssen MF, Whitehurst DGT, Mulhern B. An exploratory analysis of the pain/discomfort dimension of the EQ-5D-5L in people living with physical and mental health conditions. EuroQol Plenary, 2020.

Fayers PMD. (2016). *Quality of Life: The Assessment, Analysis and Reporting of Patient-reported Outcomes. 3rd Edition*. London: Wiley Blackwell.

Feng YS, Jiang R, Pickard AS, Kohlmann T. (2022). Combining EQ-5D-5L items into a level summary score: demonstrating feasibility using non-parametric item response theory using an international dataset. *Quality of Life Research*, 31(1): 11-23.

Feng YS, Kohlmann T, Janssen MF, Buchholz I. (2021). Psychometric properties of the EQ-5D-5L: a systematic review of the literature.

Feng YS, Kohlmann T, Peasgood T, Engel L, Mulhern B, Brazier JE, Pickard AS. (2021). Scoring the EQ-HWB and the EQ-HWB-S: Can We Do It Without Value Sets? A Non-Parametric Item Response Theory Analysis of Three Datasets. *EuroQol Virtual Plenary*

Finch AP, Brazier JE, Mukuria C. (2018). What is the evidence for the performance of generic preference-based measures? A systematic overview of reviews. *European Journal of Health Economics, 19(4):* 557-70.

Finch A, Mulhern B. (under review). Where do measures of health and wellbeing fit within a wider measurement framework? Implications for the measurement of Quality of Life and the identification of bolt-ons. *Social Science and Medicine.*

Geraerds A, Bonsel G, Janssen MF, Finch A, Polinder S, Haagsma J. (2021). Methods Used to Identify, Test, and Assess Impact on Preferences of Bolt-Ons: A Systematic Review. *Value in Health*, 24(5): 901-16.
Goosens LMA et al (2019). The Fold-in, Fold-out Design for DCE Choice Tasks: Application to Burden of Disease. *Medical Decision Making*, 39(4): 450-460.
Longworth L, Yang Y, Young T, Mulhern B, Hernandez-Alava M, Mukuria C, Rowen D, Tosh J, Tsuchiya A, Evans P. (2014). Use of generic and condition specific measures of health-related quality of life in NICE decision making: systematic review, statistical modelling and survey. *Health Technology Assessment, 18:9.*
Mokkink LB, Terwee CB, Patrick DL, et al. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol; 63(7): 737-745.
doi:10.1016/j.jclinepi.2010.02.00
Mulhern B, Sampson C, Haywood P, Addo R, Page K, Shah K, Herdman M, Janssen B et al (2022). Criteria for developing, assessing and selecting bolt-on dimensions for EQ-5D. *Quality of Life Research*.
Oppe M, Weijnen TJG, de Charro FT. (2005). Development of a questionnaire to assess the quality of care in Dutch dialysis centers from the patient's perspective. *Expert Rev Pharmacoecon Outcomes Res, 5:* 255-65.
Reeve BB, Hays R, Bjorner JB, Cook K, Crane P, Teresi J, et al. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*, 45(5 Suppl1): S22-31.
Rowen D, Zouraq IA, Chevrou-Severac H, van Hout B. (2017). International regulations and recommendations for utility data for health technology assessment. *Pharmacoeconomics, 35 (Suppl. 1):* S11–S19.
Shah K, Mulhern B, Longworth L, Janssen MF. (2017). Views of the UK General Public on Important Aspects of Health Not Captured by EQ-5D. *The Patient, 10(6):* 701-09.
Stone M. (1979). Comments on model selection criteria of Akaike and Schwarz. (Series B). *Journal of the Royal Statistical Society,* 41: 276-78.
Swinburn P, Lloyd A, Boye K, Edson-Heredia E, Bowman L, Janssen MF. (2013). Development of a disease-specific version of the EQ-5D-5L for use in patients suffering from psoriasis: lessons learned from a feasibility study in the UK. *Value in Health*, 16(8): 1156-62.
Tsuchiya A, Bansback N, Hole AR, Mulhern B. (2019). Manipulating the Five Dimensions of the EuroQol Instrument: The Effects on Self-Reporting Actual Health and Valuing Hypothetical Health States. *Medical Decision Making*, 39(4):379-392
Ware JE, Sherbourne CD. (1992). The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual Framework and Item Selection. *Medical Care*, 30: 473–483.
Wisloff T, Hagen G, Hamidi V, Movik E, Klemp M, Abel Olsen J. (2014). Estimating QALY gains in applied studies: a review of cost-utility analyses published in 2010. *Pharmacoeconomics, 32(4)*: 367-75.

Figure 1: The EQ-5D as a multi-level instrument

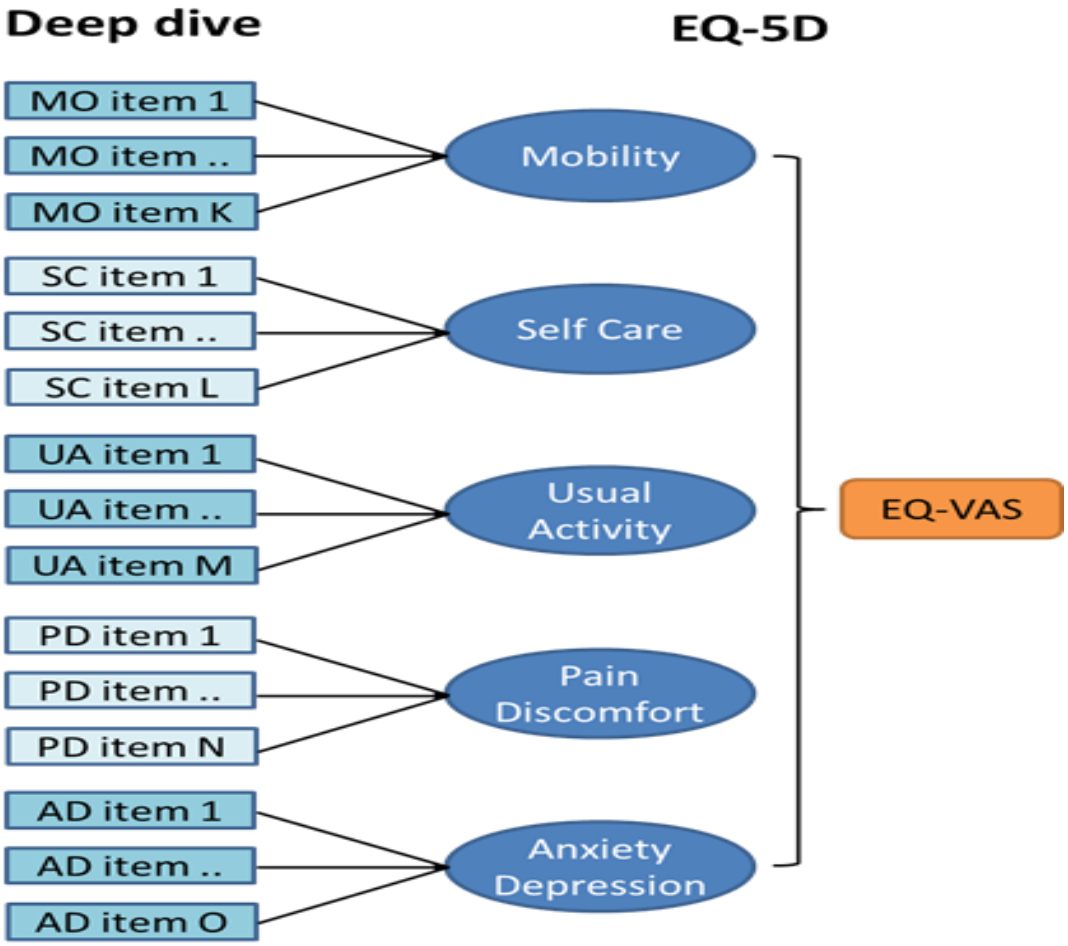

Figure 2: Structure of the Quality of Care in Dialysis Centres Questionnaire (DCDQ)

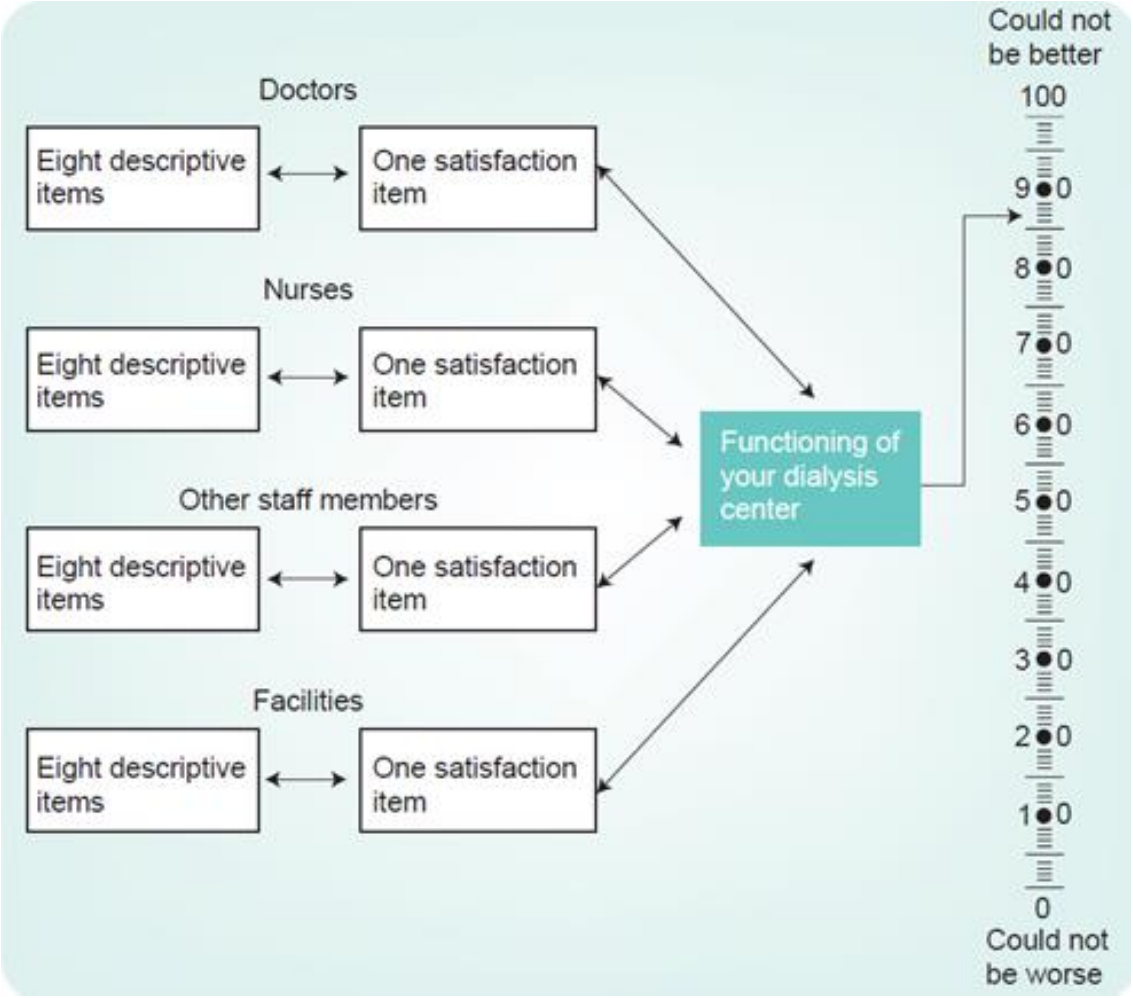

Figure 3: The AQOL-8D descriptive system structure