Value set redundancy: How should we judge whether HRQoL values remain 'fit for purpose'?

Richard Norman[1], Bram Roudijk[2], Marcel Jonker[3], Elly Stolk[2,3], Saskia Knies[4], Raoh-Fang Pwu[5], Ciaran O'Neill[6], Kirsten Howard[7], Nancy Devlin[8]

1. School of Population Health, Curtin University, Australia 2. EuroQol Research Foundation, Rotterdam, The Netherlands 3. Erasmus School of Health Policy & Management, Erasmus University Rotterdam, Rotterdam, The Netherlands 4. Zorginstituut Nederland, Diemen, The Netherlands 5. Data Science Center, Fu Jen Catholic University, New Taipei City, Taiwan 6. School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Northern Ireland 7. Menzies Centre for Health Policy and Economics, Faculty of Medicine and Health, University of Sydney, Australia 8. Melbourne School of Population and Global Health, University of Melbourne, Australia.

Abstract

Objectives: Providing policy makers with value sets is a crucial part of EuroQol's work and enables our instruments to be widely used in HTA and other settings. However, given the long history of EuroQol's work, and the number of existing valuation studies, many of our value sets are now old, based on potentially outdated valuation methodology, and in populations which no longer represent the contemporary population. Assuming EuroQol wants a role in endorsing approved value sets - or at least a way of identifying priorities for investing in new value sets - having a clear strategy for identification and mitigation of redundancy is important to ensure policy makers retain confidence in their country-specific value sets.

Methods Through discussion and iteration with our international authorship team, we have constructed a taxonomy of redundancy. We have explored how the different types of redundancy might be identified, and how the EuroQol group might work with local policy makers to address redundancy, and therefore ensure our instruments remain relevant for use.

Results The taxonomy of redundancy consists of four main areas, based on both redundancy and obsolescence. These are that the value set no longer aligns with current normative HTA requirements; that the methods used to generate it are no longer considered robust or adequately close to best practice; that the population composition has moved too far from the population in which the original value set was derived; and that, even after controlling for population differences, preferences are likely to have changed since original data collection. Through identification of the type of redundancy that applies in a particular setting, we then suggest a range of possible solutions to each, ranging from recommending particular sensitivity analyses, through reweighting of existing data to better account for population differences, to collecting new data for an updated value set.

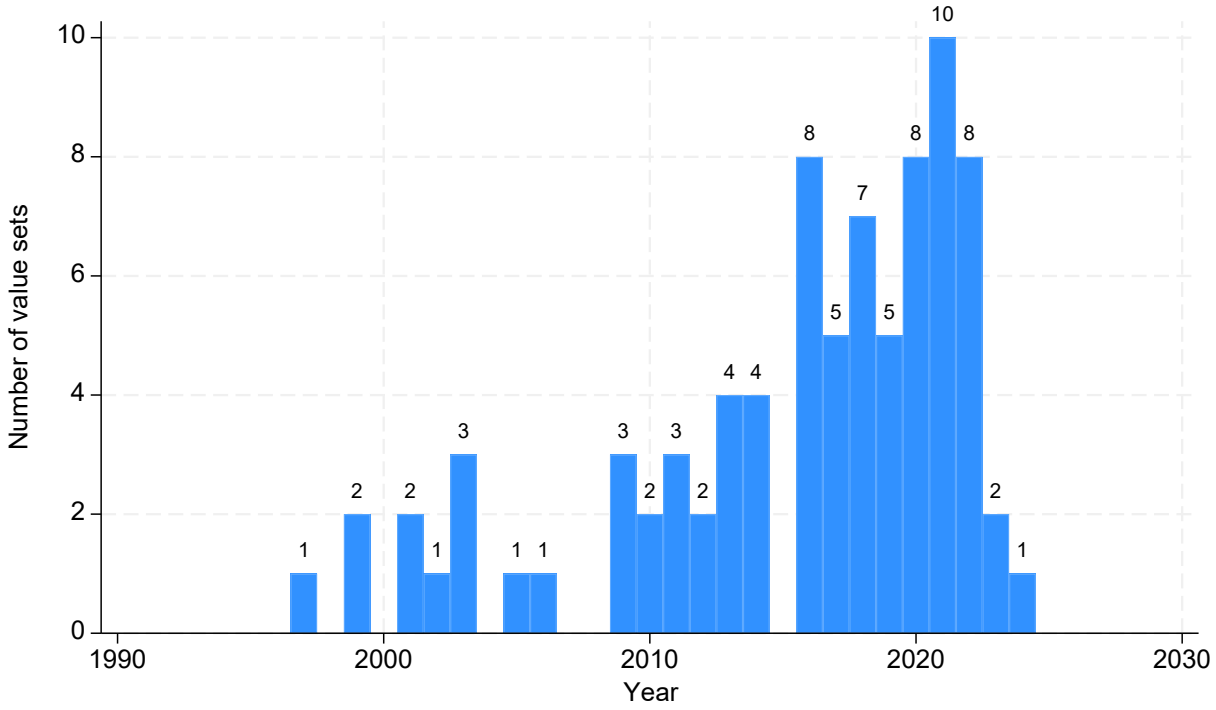
Conclusion: Redundancy of existing value sets is driven by more than just time since data collection and is often a matter of judgment rather than based on a clear definition. Working closely with local policy makers, in manner appropriate to the local environment, to discuss the ongoing appropriateness of existing value sets is an important part of EuroQol's ongoing role, and includes the consideration of updating value sets in contemporary populations using current best-practice methods. However, the benefits of updating value sets has to be balanced against the desire of policy makers for consistency in their local decision-making processes.

1. Introduction

Value sets for health-related quality of life (HRQoL) instruments are widely used to support decisions about health and health care in a range of settings and applications. This use is based on the premise that improving health is a core function of the health and health care sector, and that HRQoL is a central component of this. These applications can be divided into two broad categories. First, value sets are used in 'quality weighting' life years in the calculation of quality-adjusted life years (QALYs) in cost utility analysis (CUA) of health care interventions. This evidence is widely used to inform health technology assessment (HTA) and other decisions concerning health care resource allocation. Second, value sets are also used as a convenient means of summarising the profile data generated from HRQoL instruments (such as EQ-5D) into a single number, for ease of statistical analysis. For example, values have been used by health care systems (e.g. the English NHS) to summarise EQ-5D data collected as part of routine outcomes measurement (PROMs), to evaluate treatment effectiveness and assess provider performance. The use of value sets in 'QALY' and 'non-QALY' applications may have different implications for the properties required of value sets (Devlin, Finch, Parkin 2022).

The development of HRQoL values for QALY estimation has a tradition dating back half a century (Spencer et al 2021). Country-specific value sets for EQ-5D instruments began to be generated from 1997 (with publication of the seminal MVH study – Dolan 1997); thus there are value sets which are now more than 20 years old. Using data from the EuroQol website, the general population value sets for the EQ-5D-3L, the EQ-5D-5L, and the EQ-5D-Y-3L are plotted in Figure 1 and illustrate some of the context behind this current paper.

Figure 1: EQ-5D Value Sets, by Year of publication



While there are instances of value sets being updated (the EQ-5D-3L in Slovenia, the EQ-5D-5L in China), most of these older value sets remain current. Where multiple value sets do exist, they tend to co-exist with limited guidance for users about their use. This gives rise to the question about the extent to which these older value sets still offer adequate input for decision making. Value sets may become outdated for various reasons and, at a certain point, deemed redundant, i.e., not useable for their original intended purpose. To date, this issue has not been explicitly addressed by stakeholders (including instrument developers, researchers, policy makers and patients), and this represents an important gap in the literature, increasingly so as time and methods progress.

The aims of this paper are to (a) discuss key issues in the definition of redundancy in value sets, (b) provide a taxonomy of the various factors that contribute to a given value set being deemed redundant, (c) consider the criteria (and related evidence requirements) which instrument developers (or users of value sets) could use to judge value set redundancy and to identify the need for updated value sets, and (d) to highlight implications of a decision to update/replace a value set, such as transitional issues for value set users and decision makers in switching between value sets with different properties.

The following paper considers these issues in the context of EQ-5D value sets only. The key reason for this is that there are more country-specific value sets for EQ-5D instruments than for any other HRQoL instrument. With over three decades of research investment in EQ-5D

valuation methods and value sets, including new methods, standardised protocols and quality assurance processes, there is a need to develop a clear rationale and process for decisions about EQ-5D instruments' value set redundancy, and this paper is a response to that need.

However, the issues discussed in this paper are also highly relevant to other HRQoL instruments. Indeed, EuroQol's ongoing investment in research on valuation methods and willingness and ability to create new value sets in a large number of countries makes it a notable exception among HRQoL instrument developers. In many other cases, generic preference-weighted instruments (of the kind widely used in QALY estimation) are accompanied by a small number of value sets (often one). Examples include HUI3, for which only one a small number of value sets exist (for example, Feeny et al 2002). Other examples include the 15D for which only one value set exists (Sintonen 1995), obtained from Finnish adults. An important recent example is PROMIS, for which one value set, PROPr, is currently available, based on the stated preferences of US adults (Dewitt et al 2018). This 'single value set' approach is common, and these value sets remain in use even when they are noted as having problematic features[1]. Therefore, the issues regarding value set redundancy we investigate in this paper are relevant to *all* HRQoL instruments accompanied by preference weights.

2. Challenges in defining redundancy in value sets

It is important to note that it is challenging to define value set redundancy, to determine who is responsible for making such a decision, and to explore the appropriate course of action resulting from a value set meeting redundancy criteria. On the first of these, to date, value set redundancy has neither been defined nor the criteria used to identify it explicitly identified. Therefore, the field exists with this uncertainty and the identification of value set redundancy has been open to interpretation and *ad hoc* judgement.

We take as our starting point that redundancy is linked to the question of whether or not a value set is considered 'valid' in a contemporary setting. There is rarely an external gold standard which can be used to judge the validity of a value set, and there is no agreed definition of what 'validity' means in the context of HRQoL values. It can be argued that it is "almost impossible" to validate HRQoL values in the way we can validate stated preferences in other applications and sectors. There are few opportunities to observe "real" choices people make about HRQoL, so we lack the kind of revealed preferences data that would allow us to check that values are meaningful representations of the preferences embodied in decisions (Devlin et al 2022).

[1] For example, the HUI3 value set has negative values for 78% of the state in its descriptive system; and the PROPr utilities that accompany PROMIS-29 have been shown to have a number of odd characteristics (Pan et al 2022).

However, we do need to advance a working definition for this work to proceed. For the purposes of this paper, we tentatively propose the following definition of HRQoL value set validity:

HRQoL value set validity concerns the extent to which any given set of values for an HRQoL descriptive system (a) are a sufficiently good representation of the average preferences of the population of interest and (b) have empirical and theoretical properties which are acceptable in the decision-making context.

The definition touches on two things. First, it considers whether the values adequately reflect the average preferences of the members of a given society or some sub-set of it deemed to be relevant on normative grounds. For example, NICE's methods guide notes that values for adult HRQoL should be obtained from adult members of the general public. Regarding the term "adequately good representation" we suggest such a definition cannot be easily made more precise. The second part of the definition of validity concerns whether the characteristics of the values are a good match with any stated requirements of decision makers and have empirical characteristics with desired properties. This includes the basic requirement, for use in QALY estimation, that values be anchored at 0 and 1 (but can lie below 0 if health states are considered to be worse than being dead) and should have interval scale properties. For example, NICE's methods guide notes that values should be 'choice based' (indicating a requirement around methods). Value sets which meet such decision-maker requirements might be considered to have 'context validity' (Bailey et al 2023). Every aspect of the research process employed to produce value sets may give rise to considerations regarding appropriateness, acceptability and whether the resulting values are 'fit for purpose,' i.e., choice sample frame, methods used to elicit stated preference, quality assurance processes applied during or after data collection, modelling approaches, and so on.

Using this working definition of the validity of HRQoL values, in the following section we identify factors which arguably *compromise* validity and which might lead to value set redundancy.

3. A taxonomy of factors which affect value set redundancy

Type 1 redundancy – the value set misaligned with normative views

We believe that a value set is redundant in a particular context if the decision-making body advocates, or moves towards, a different normative basis for deriving value sets. For example, given the increasing focus on HTA incorporating patients' perspectives, there may be a shift toward seeking patients' values for HRQoL. Similarly, in the valuation of child health, there is increasing interest from stakeholders (e.g. in the US, UK and elsewhere) in HRQoL values

reflecting children's own views about their health. In both case, value sets based exclusively on the stated preferences of the adult general public may become less relevant as the sole basis for generating evidence, and this would typically trigger further valuation work to develop results using the preferred normative approach. On a related point, it may be that this kind of redundancy can apply to certain kinds of analysis within the same jurisdiction. For example, value sets obtained from general population preferences may align well with what decision-making bodies need, but might be less relevant and appropriate as a means of summarising patients' data in the context of PROMs programmes (e.g. where the goal is to measure the performance of procedures or providers in improving patient health).

Similarly, we argue that a value set becomes redundant if the relevant decision-making body moves away from the use of a particular instrument to support HTA or other decision making. If a value set is for an instrument that is no longer recommended by a particular governmental body, then the value set is itself in a sense redundant for that purpose, but would not require a further valuation survey (but might cause the need to begin valuation of health states described using other replacement instruments).

In an extreme case, the decision-making body might move away from the QALY metric as a central part of their processes. Such a move could require a complete reconsideration of how HRQoL is integrated into the decision-making process. This would depend on the selected alternative; for example, GRACE (Lakdawalla and Phelps 2021) continues to require assessments of HRQoL. But if the alternative paradigm did not use such measures, then the value set would be redundant, but this redundancy would not trigger a new valuation project.

Type 2 redundancy – methods used have become outdated and/or unreliable.

A wide range of changes in valuation methods have occurred in recent years, including the type of stated preference tasks, mode of task administration, quality control processes, data analysis and modelling methods. These changes arise for a variety of reasons. Some arise as pragmatic responses to circumstances, such as the shift to online interviews as a result of the pandemic. Others arise from changes in underlying theoretical emphasis, such as recent discussions over the role of time preference in trade-offs between quality of duration of life, leading to interest in non-linear DCE methods. Such changes can be broken down into those supported by strong scientific evidence around methods superiority, and those that represent a change in approach *preferred* by methodologists (e.g. because they prefer one kind of underlying theory to another e.g. random utility theory vs. utility under uncertainty), although the dichotomy will often be much less clear, with changes reflecting elements of both.

If original data analysis can be updated to (for example) run a different model, or exclude data no longer considered reliable, then analysis can simply be re-run. Hence the value set

might be redundant, but amenable to updating and thus not causing the need for new valuation data to be collected.

However, if the original data analysis that generated the value set cannot be updated, then it may be worth exploring whether the magnitude of the effect be estimated, and hence inform a decision whether or not to rely on the older value set, or to conduct new valuation work. One option here would be to run a small methodological study using previous and new methods to quantify the difference. If it is demonstrated to exist and to be of an adequately large size to matter (however that might be defined), then that would then trigger the conduct of larger valuation study using updated methods.

Type 3 redundancy – population has changed since the original valuation work

Over time, populations change both in their composition (type 3a) and their preferences (type 3b). With respect to 3a, even if the average preferences of any one sub-group of society (e.g. defined, by age, culture or any other factor(s)) remained unchanged through time, a change in the composition of the population (e.g. arising through an ageing population, or through patterns of immigration), could change the overall average 'societal' preferences. With respect to 3b, changes in health state preferences might plausibly arise through time as a result of changing societal expectations about HRQoL; greater awareness of types of health problems (e.g. mental health); and as a result of relevant issues being debated at a societal level (e.g. experiences relating to the COVID pandemic, euthanasia, end-of-life care, or abortion).

The kinds of changes in 3a and 3b might be addressed in quite different ways, with the latter relatively more likely to trigger new valuation data collection. Regarding population compositional change (type 3a), existing data can in principle be reweighted to explore the magnitude of the effect, and to potentially develop an updated value set.

Regarding preference change independent of population composition (type 3b), it may be that we need to monitor preferences using a standard, low-cost survey (e.g. using latent scale DCE), which can, if result indicate a change, trigger a fuller new valuation study. If underlying preferences have changed, then that represents evidence that the original value set has moved towards redundancy. However, it is also important to ensure that any change has re-stabilised around new norms, potentially through a series of low-cost surveys.

Type 4 redundancy – the instrument has changed and now the value sets is not an exact match for the descriptive system

The development of value sets occurs subsequently to considerable instrument development and refinement. While instrument developers will tend to finalise an instrument before

valuation commences, it may be that evidence accrues around the appropriateness of the instrument subsequent to valuation work being disseminated. If this prompts the developer to update the instrument, then it may be that any valuation work done on the outdated version of the instrument is similarly redundant. The question to be addressed by the group responsible for identification of value set redundancy is whether the change in wording is likely to produce different values if the same valuation study were conducted using updated wording.

4. What evidence is needed to test for each type of redundancy and what solutions are suggested?

Table 1 summarises the evidence required to test for each type of redundancy, and the likely solution if redundancy is identified.

Table 1: What evidence is needed on each type on redundancy, and what actions are possible in each case?

A. Redundancy Type	B. What evidence would be required to test for redundancy?	C. What solutions are possible if there is evidence of redundancy?
1. The value set no longer aligns with current normative HTA requirements	None as it is driven by the underlying methodological guidelines of the HTA body	Development of new value set better aligned with guidelines
2. Methods used have become outdated and/or unreliable	Evidence that the value set is likely to change due to changing methods	If data can be re-analysed using contemporary methods, this is optimal. If not, retain current value set if changes are shown to be modest. Or, if not, development of new value set using gold standard methodology
3a. Change in average preferences, due to changes in population composition	Resampling of original data to explore whether there is a significant change in mean preferences	Assuming the appropriate population characteristic data was collected, re-weighting of existing responses to better account for new population composition
3b Change in average preferences, due to changes in society's preferences	Indication of changing attitudes, such as qualitative work, or small quantitative	Development of new value set using gold standard methodology, and contemporary sample

	work exploring preferences	
4 The instrument has changed and now the value sets is not an exact match for the descriptive system	Small valuation study using original and updated wording	Development of new value set using updated wording

It is important to note that the solution may not always require new data collection. It may be possible to re-analyse existing data in new ways, using improved modelling methods; impose higher standards of quality control *ex post* by excluding data; or to re-weight data to address changes in the composition of the population.

However, in some cases it will be necessary to undertake a new value set study to replace the redundant one. Given the cost (both financial and in a broader sense described in Section 5) of undertaking such studies, there should be clear evidence on agreed criteria that these efforts are warranted. Further – as we discuss in the next section – new value sets are not just costly to produce, but also impose costs in terms of implementation which need to be taken into account when deciding to denote a value set as redundant.

5. The cost of transitions to new value sets and implications for judging redundancy

For Type 1 redundancy, the case for value set redundancy is normally clear. However, for types 2 and 3 redundancy, there is a balance between the advantage of a more contemporary value set using current gold-standard methods, and the acceptability of their update to decision-makers. Updating value sets has the advantage of better reflecting the values of the community in which decisions are being made (either by administration in a more contemporary sample, or in using methods which we as a field believe to be more rigorous than what was the gold standard previously). However, this updating process comes at a cost. First, deriving value sets is expensive and draws resources away from other research. This argument may benefit from the development of an EVPI-type framework, and it should also be noted that the cost of resource misallocation based on redundant HRQoL values can be significant. Second, having a new value set requires good stakeholder engagement to ensure there is comfort with switching to it in preference to the widely used existing value set. Change has to be well justified given the potential for gaming in HTA where multiple competing value sets are available; it may be that commissioning of new value sets has to operate in tandem with a process of actively decommissioning of older value sets. However, given we are operating in an environment without external validation of values, can we equivocally say the older value set is inferior and hence should be decommissioned? For

Type 1 redundancy, that is easier, but our expectation is that Types 2 and 3 redundancy will be more common - and will eventually apply to all value sets.

A further point is how HTA processes should use new value sets. A new value set will change how quality of life and length of life are valued against one another, and also the relative importance of different aspects of quality of life. Therefore, QALYs estimated under different value sets are not necessarily comparable. If a value set is replaced, should that change how decision-makers consider ICERs in an HTA context? And therefore, how do decision makers ensure that decisions with a new value set remain consistent with older decisions? Is this best achieved through HTA bodies recommending value sets, but requesting standard sensitivity analyses using competing value sets? To our knowledge, there is currently no clear guidance on this provided by any HTA body, but there is good theoretical and empirical data suggesting it requires consideration as value set selection can change results significantly.

If a new value set replaces a previous one, does that then cause a problem in terms of the need to reappraise historical decisions? If we assume that a new value set is correct, and the older one is not, and that switching between value sets moves interventions across some cost-effectiveness threshold, should policy makers then reverse decisions in light of new evidence? It is highly unlikely that positive recommendations would be reversed if the value set were to change the implied ICER, but it is certainly plausible that sponsors would ask for reconsideration of evidence if previously rejected interventions become more cost-effective when a new value set is applied to the data that they previously presented. This asymmetry poses a problem through recommendation of interventions with poor cost-effectiveness data.

Conclusion

In this paper, we have tentatively identified an emerging problem for developers of HRQoL instruments with accompanying value sets, such as EuroQol. As time and methods advance, the bedrock of applied valuation research naturally becomes increasingly unreliable, and as a field, we need to consider how to approach this challenge. Here, we have presented a framework for describing and addressing value set redundancy, but have left questions unanswered. Some questions - such as those around how large a difference in expected values warrants new valuation work, we believe are best addressed on a case-by-case basis, and at worst, may be unanswerable as we do not know how big a difference is 'too big to ignore'. However, some other questions, such as the value of adjusting existing data for different population composition, and the best way to engage with policymakers around this issue, are fruitful avenues for ongoing research, and something we would be keen to see

taken on by the field more generally to help keep our value sets fit for purpose and reflective of broader societal views.

Possible questions for discussion

1. Should the EuroQol Group have a role in judging whether value sets are 'redundant', or is it a case of *caveat emptor* for local users and decision makers?
2. To what extent is decision maker resistance to 'new' values, and the potential inconsistencies between 'old' and 'new' QALY estimates arising from applying different value sets to HRQoL data, something we should take into account when thinking about redundancy? Is there more we can do to help decision makers handle transitions between old and new value sets?
3. Other generic instrument developers seem content to rarely, or never, update their preference weights. Does EuroQol obsess too much over value sets? Or is the availability of updated value sets across multiple countries a key strength we should highlight more?

Funding: Work on this paper was supported by EuroQol Research Foundation grant EQ 1578-RA. Views expressed in this paper are those of the authors, and are not necessarily those of the EuroQol Research Foundation.

References

Bailey C, Howell M, Raghunandan R, Howard K, Mulhern B, Petrou S, Rowen D, Salisbury A, Lancsar E, Devlin N. (2024) The RETRIEVE Checklist for Studies Reporting the Elicitation of Stated Preferences for Child Health-Related Quality of Life. *PharmacoEconomics* https://doi.org/10.1007/s40273-023-01333-zDevlin, N., Finch, A.P., Parkin, D. (2022). Guidance to Users of EQ-5D-5L Value Sets. In: Devlin, N., Roudijk, B., Ludwig, K. (eds) Value Sets for EQ-5D-5L. Springer. https://doi.org/10.1007/978-3-030-89289-0_5
Devlin NJ. Valuing Child Health Isn't Child's Play. Value Health. 2022 Jul;25(7):1087-1089. doi: 10.1016/j.jval.2022.05.009. Epub 2022 Jun 3. PMID: 35667949.
Dewitt B, Feeny D, Fischhoff B, Cella D, Hays RD, Hess R, et al. Estimation of a preference-based summary score for the patient-reported outcomes measurement information system: the PROMIS((R))-preference (PROPr) scoring system. Med Decis Mak. 2018;38(6):683–98
Dolan P. Modeling valuations for EuroQol health states. Med Care. 1997 Nov;35(11):1095-108. doi: 10.1097/00005650-199711000-00002. PMID: 9366889.
Feeny, D. (2002). The utility approach to assessing population health. Summary measures of population health: Concepts, ethics, measurement and applications. J. S. C. Murray, C. Mathers, & A. Lopez. Geneva, World Health Organisation: 515-528.

Lakdawalla DN, Phelps CE. Health Technology Assessment With Diminishing Returns to Health: The Generalized Risk-Adjusted Cost-Effectiveness (GRACE) Approach. Value Health. 2021 Feb;24(2):244-249. doi: 10.1016/j.jval.2020.10.003. Epub 2021 Jan 12. PMID: 33518031

Pan, T., Mulhern, B., Viney, R. *et al.* A Comparison of PROPr and EQ-5D-5L Value Sets. *PharmacoEconomics* **40**, 297–307 (2022). https://doi.org/10.1007/s40273-021-01109-3

Sintonen H. (1995) The 15-d Measure of Health Related Quality of Life. II Feasibility, Reliability and Validity of its Valuation System. Centre for Health Program Evaluation Working Paper 42. Monash Centre for Health Economics.

Spencer A, Rivero-Arias O, Wong R, Tsuchiya A, Bleichrodt H, Edwards RT, Norman R, Lloyd A, Clarke P. (2021) The QALY at 50: One story many voices. Soc Sci Med. 2022 Mar;296:114653. doi: 10.1016/j.socscimed.2021.114653. Epub 2021 Dec 11. PMID: 35184921.