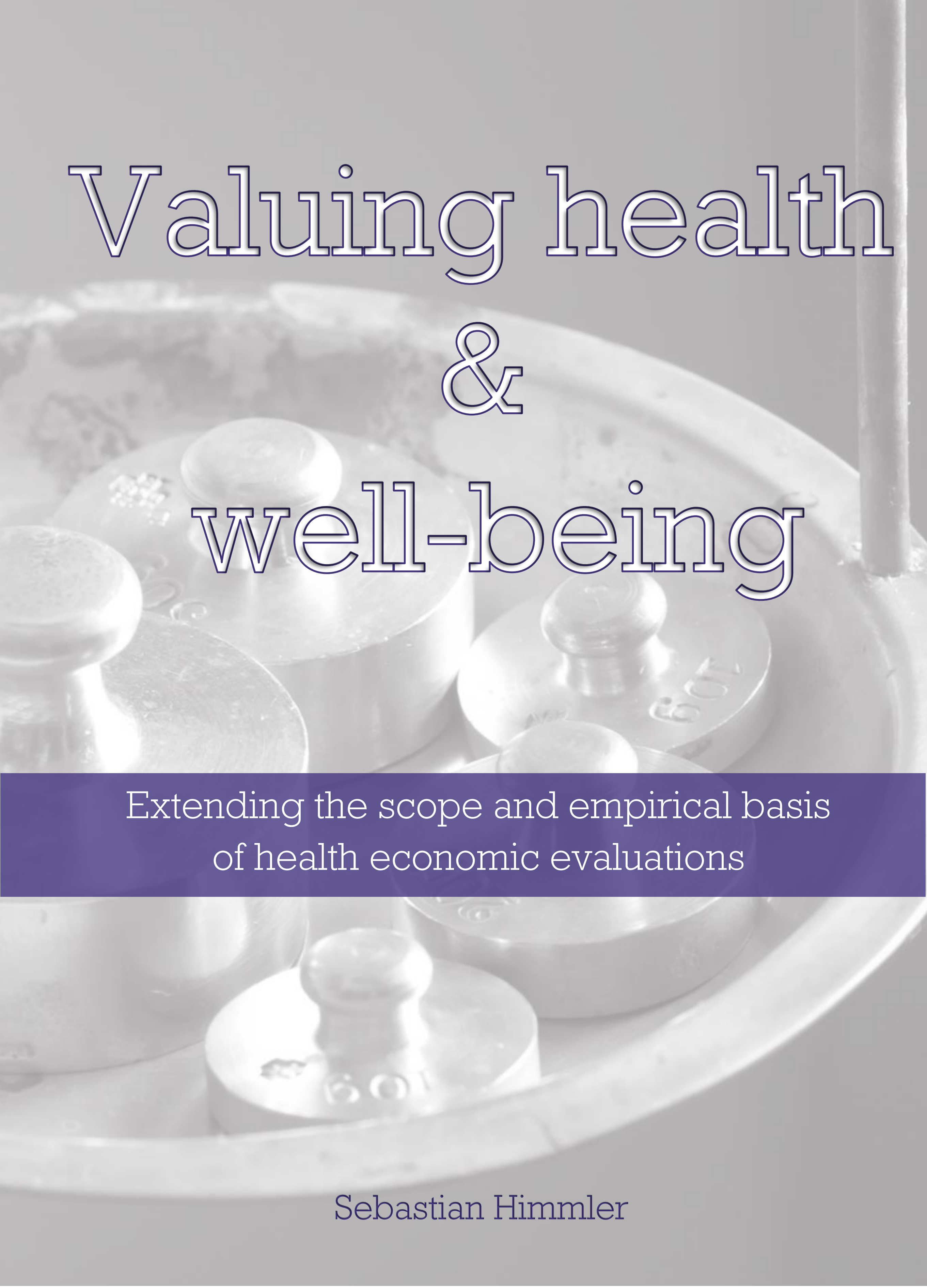


Valuing health & well-being

Extending the scope and empirical basis of health economic evaluations

Sebastian Himmler

Valuing health and well-being

Extending the scope and empirical basis of health economic evaluations

Sebastian Friederich Wolfgang Himmler

ISBN: 978-94-6416-702-3

Layout: Sebastian Himmler
Printing: Ridderprint, www.ridderprint.nl

Valuing Health and Well-being
Extending the scope and empirical basis of health economic evaluations

Waarderen van gezondheid en welzijn
Uitbreiden van de scope en empirische basis van gezondheidseconomische evaluaties

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof. dr. A.L. Bredenoord

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on
Friday 11 November 2022
at 10:30 hrs

by

Sebastian Friederich Wolfgang Himmler
born in Sulzbach-Rosenberg, Germany

Erasmus University Rotterdam

Doctoral committee

Promotors:	prof. dr. W.B.F. Brouwer
	prof. dr. N.J.A. van Exel
Other members:	prof. dr. E. de Bekker-Grob
	prof. dr. J. Schreyögg
	prof. dr. C. Dirksen

Table of contents

1

Introduction

Cool heads and warm hearts*

As health economists, we spend our time looking at 'quantities' like quality of life, life expectancy, numbers of deaths, or costs. We weigh up gains and losses in health and look at the associated costs, allegedly with an analytical, cool mind. However, at the same time, we try not to forget that behind every number, there is a potential life, a person with needs, wishes and preferences. It is not only the pursuit of knowledge that motivates us in this profession. It is also the compassion with our fellow human beings and the conviction that with health economic evidence, societies can make better health care decisions. Decisions informed in that way have the potential to maximise health and well-being in the population given a limited budget for health care. Consequently, as health economists we should have cool heads when creating our models and doing our calculations, but warm hearts when conceptualising what our models should include (with room for aspects like equality and equity), interpreting our findings, and formulating our recommendations.

Some health economists sit in committees where decisions on the reimbursement of new, potentially life-saving interventions are made. The health economist's heart may be with the patients that describe their misery due to a certain disease, potentially eased by a new expensive intervention (e.g., medical treatment). At the same time, the health economist's head must be with the unknown patients that are not in the room. A decision in favour of reimbursing an expensive intervention for one group of patients may have adverse consequences for other groups. For example, due to a limited budget, the interventions they need might then not be funded, or no longer be funded. This leaves the question; how should such difficult decisions be made? Our task as health economists is to provide insight into the costs and benefits of healthcare interventions to society. The following section will give a broad introduction to how such health economic evaluations of interventions work and important considerations that play a role.

Two aspects are worth noting at this stage. First, most health economists would agree that decisions based on systematically provided evidence should always be preferred to decisions made based on 'gut feelings' or past experiences. A systematic approach increases accountability and transparency in decision-making, if not also welfare.[1] Second, they would also acknowledge that the health economic perspective is always only one of the possible perspectives one can adopt in a certain decision context. Other perspectives will always also have a weight in health care decision making, not in the last place because not everything of value for decision making is easily quantified and captured in health economic evaluations.

[*] This heading is inspired by the first chapter of "A little history of economics" from Niall Kishtainy, who used this quote from the famous economist Alfred Marshall describing the ideal qualities of an economist.

Health economic evaluations

Public health care systems are collectively funded. As such the resources available for health care are in direct competition with private income and expenditures in other public sectors like social security, infrastructure, education, or defence. More resources could be made available for health care by shifting resources between budgets or increasing the budget by raising taxes or contributions, but, in the end, the available resources for health care are essentially limited. At the same time, health care costs are expected to keep increasing due to three factors: the ageing of the populations, technological advances, and rising expectations.[2] Therefore, regulatory health care bodies like ZIN in the Netherlands, NICE in the UK, or G-BA in Germany, have to decide whether new medical interventions should be made available in the public health care benefits (or insurance) package. The goal of health economic evaluations is to inform such decisions by providing information about the incremental costs and benefits of a new intervention compared to current standard care.[3] In essence, health economic evaluations attempt to systematically analyse the value of interventions using explicit and commonly agreed upon assumptions, which helps to increase the quality, transparency, and accountability of health care decisions.

In the health care context, the most prominent type of health economic evaluation is cost-effectiveness analysis (CEA). In most cases, CEA entails calculating differences in health outcomes and costs between a new intervention and an existing intervention. Dividing differences in costs by differences in health outcomes builds the incremental cost-effectiveness ratio or ICER, which represents the additional cost per unit of outcome. Health outcomes can take many different forms in CEA depending on the disease or intervention under consideration, from, e.g., reduced blood pressure to more fundamental outcomes like life years saved. Comparing results from different CEAs can be challenging if different disease-specific outcomes are used.[3] Therefore, the main alternative to CEA in health care, and one of the focal points of this dissertation, is cost-utility analysis (CUA), where the health outcome is captured in terms of Quality-Adjusted Life Years (QALYs).

QALYs are a generic outcome measure combining quality of life (or more specifically, health-related quality of life) and length of life.[1] Health-related quality of life is measured using generic, multi-dimensional, instruments like the EQ-5D[4] or the SF-6D.[5] The former, for example, has the following dimensions: Mobility, self-care, usual activities, pain/discomfort, and anxiety/depression.[6] Oftentimes using a survey, patients indicate how they are doing in each of these dimensions. The different levels of health selected in each of the distinct dimensions then describe the health state of the patient. This type of data is predominantly collected alongside clinical trials of new interventions. Population preferences are used to compute a utility score for all possible health states on a scale between 0 (the state of being dead) and 1 (the state of full or perfect health). A short example: An individual indicates to have some problems in the mobility dimension of the EQ-5D, while having no problems in the other four dimensions. The health utility of this health state is 1 (full health) minus the weight for having some problems in the mobility dimension. This weight was estimated to be -

0.04 for the Dutch population.[7] Consequently, the utility of this health state is 0.96. Such quality-of-life values are then combined with length of life (in years) to build QALYs (i.e., 10 years in this health states equals to 9.6 QALYs). How the QALY concept works in the context of comparing two health interventions is illustrated in Figure 1.

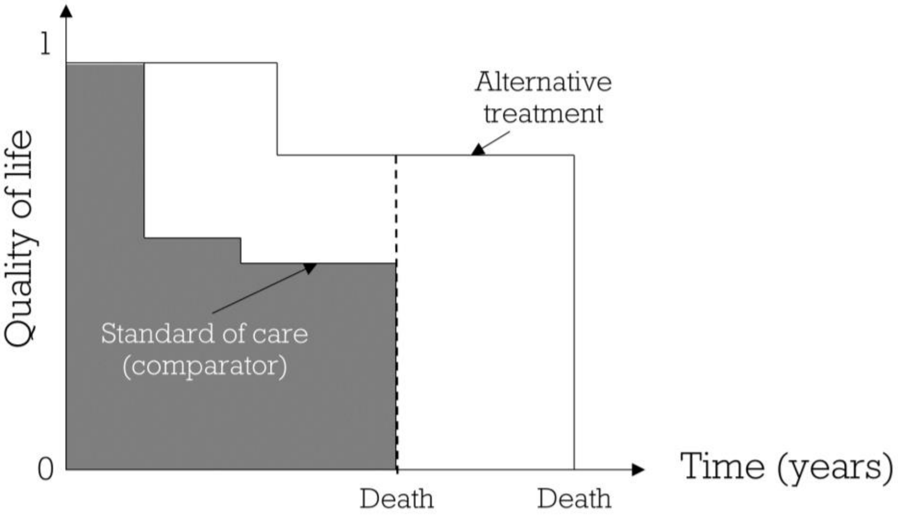


Figure 1: Depiction of incremental health gain in quality-adjusted life years (QALYs). Own illustration after Gold et al. (1996).[8]

The vertical axis represents the health-related quality of life or health utility ranging from 0 (dead) to 1 (full health). The horizontal axis depicts the remaining life expectancy of the patient (group). The grey area then represents the expected quality and quantity of life associated with being treated for a certain disease with standard care. The white area represents the expected additional QALYs gained due to the alternative intervention. The white area left of the dotted line represents the QALY gain through quality-of-life improvements from a new intervention, the area right of the line the QALY gain through extending quantity of life, i.e., increasing life expectancy.

The above-mentioned cost-effectiveness ratio, or ICER, then means dividing differences in costs between interventions, by differences in QALYs. For example, a new intervention costs €10,000 per patient more than current standard care. At the same time, the new intervention increases length of life to the extent of 2 QALYs. In this example, the ICER would be 5,000 €/QALY. But how should ICER values be interpreted to form policy recommendations? For this purpose, Figure 2 spans the widely used cost-effectiveness plane, dividing potential ICER results into four quadrants (A-D). Here, standard care is located at the interception of the two axes, and compared to two alternative interventions, medical treatment I and medical treatment II. The horizontal axis depicts the incremental QALYs, i.e., the difference in health provided by the new treatment compared to standard care. The vertical axis represents the incremental costs associated with each alternative treatment compared to stand care. Depending on whether alternative treatments produce more or less QALYs and are more or less costly than the comparator, treatments may lie in either of the four quadrants. What do these imply in terms of policy recommendations? In quadrant B, treatments would be more costly and produce less QALYs than standard care, leading to the recommendation that the intervention should not be implemented. In quadrant D, treatments would cost less and produce more QALYs, which would result in an unequivocal positive recommendation.

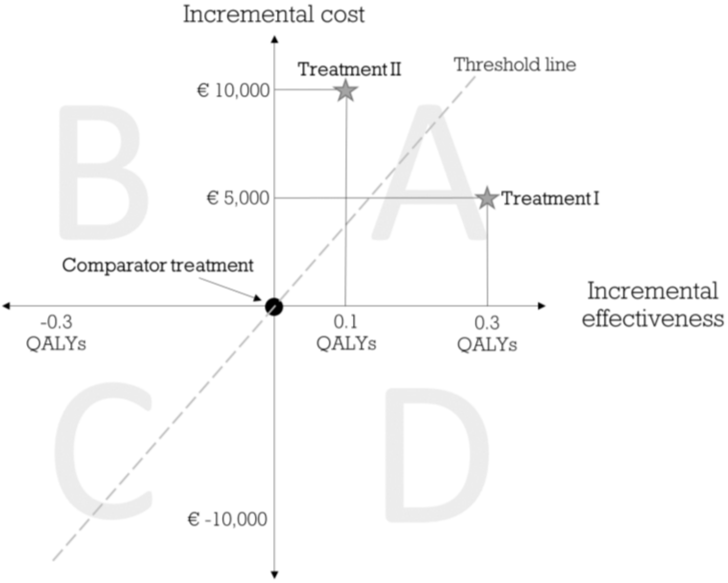


Figure 2: The cost-effectiveness plane. Own illustration after Drummond et al. (2015)[1]

Treatments in quadrant C are rarely assessed, as effectiveness is often a dominant decision criterion, i.e., the acceptance of less effective new technologies is typically low in the medical field also when they are much cheaper than standard care. Most new treatments lie in quadrant A. These provide a larger health benefit but at the same time are more costly than standard care. The question then becomes: should such an intervention be implemented in the benefits package? The answer is 'it depends'. Everyone probably agrees that paying €100 for one year in full health, i.e., 1 QALY, would be money well spent. But what about paying €100,000 for one QALY? As such, to be able to interpret ICERs in quadrant A, some sort of 'threshold' or value of a QALY is needed. Above such a threshold, a treatment should not be considered worthwhile, offering value for money, or being cost-effective anymore. Ignoring any other considerations for now, it boils down to the question what an acceptable ratio of incremental costs per QALY is. In other words: how much is society willing to pay for an additional unit of health. This decision rule, in its most general sense from a societal welfare perspective, can be formulated as in equation 1:

$$\frac{\Delta C}{\Delta Q} < v_Q \tag{1}$$

The ratio of incremental costs ΔC and incremental QALYs, ΔQ, should be lower than some threshold v_Q (the nature of which will be outlined in the following section). If this is the case, an intervention is considered cost-effective and the recommendation would be to fund the intervention.[9] Under certain assumptions, this decision rule implies the maximisation of the number of (gained) QALYs for a given budget.

After this introduction to the general approach of health economic evaluations, a short side note on some normative and ethical aspects of this 'decision rule' is due. Some people would categorically object to withholding effective interventions due to monetary or budgetary considerations, or costs in general. One important clarification is that economic evaluations support *collective* funding decisions at the health care system level. This is detached from the clinical level and individual patients, and not related to bed-side rationing. Second, for health economists the focus on 'costs' is not a means in itself and in no way a call for health care budget cuts. If one accepts the notion that health care resources are limited and cannot be endlessly extended, the question for economists is how these resources can be spent in the most efficient way: Every Euro spent on a certain intervention for one patient group could perhaps also have been put to alternative use for the same or another group of patients. The health gain from this potential alternative use of the budget is now not realised. Therefore, in a health economic evaluation the costs represent the sacrifices imposed on alternative uses of these resources (within or outside the health care sector). This was once formulated by Williams (1992), in the early days of health economic evaluations, as follows:[10,11]

"Anyone who says that no account should be paid to costs [in medical practice] is really saying that no account should be paid to the sacrifices imposed on others. I cannot see on what ethical grounds you can ignore the adverse consequences of your actions on other people."

At the same time, the use of QALYs in CUA do raise certain questions with an ethical dimension.[12] Of special importance are equity considerations. At their core, these relate to whether, under certain circumstances, one should give some QALYs more weight than others, or even deviate from the general notion of maximising QALYs. Equity considerations have, for instance, been formulated in the context of the age of beneficiaries, the severity of the disease, the size of the patient population, or the rarity of the disease.[13] How to incorporate equity considerations into health economic evaluations is discussed in detail elsewhere.[14,15]

Reflections on specification and empirical basis of health economic evaluations

As outlined in equation (1), the decision rule in cost-utility analysis is that an intervention is cost-effective if the ratio of incremental costs and incremental benefits (measured in QALYs) is below the cost-effectiveness threshold v_Q. Although this appears to be a simplistic framework, the exact specification and measurement of ΔC, ΔQ and v_Q is not straightforward, reason for long-standing debates, and often depends on the context. For instance, the natural starting point for specifying ΔC is resources used within the health care sector. But it could also include costs from the informal sector, costs occurring in other sectors (e.g., education), productivity costs, or future (unrelated) medical costs.[16,17] The focus of this dissertation, however, lies on the scope and measurement of the two remaining factors in equation 1: Q, the benefit dimension of cost-utility analysis, and v_Q, the cost-effectiveness threshold.

The scope of Q

The benefit dimension of cost-utility analysis, Q, is predominantly measured using QALYs, which are calculated based on instruments for measuring health-related quality of life like the EQ-5D. On first sight, it seems to be a logical approach that health care interventions should be evaluated using some type of approximation of health as benefit dimension. However, in recent years, researchers have started to question this paradigm. For instance, it has been pointed out that the maximisation of health may not be in alignment with society's values towards health care and its function.[18] Furthermore, health improvement may not be the appropriate objective in all areas of health care. If you take palliative care, for instance, restoring health is often not possible anymore. To a lesser degree, this is also the case for many interventions in elderly care, arguably one of the largest areas within the health care sector. Mental health and integrated social care are additional areas, where traditional measures of health may fall short of capturing the benefits of the care provided and of new interventions.[19,20] In the mentioned areas, interventions rather aim to maintain or increase the broader well-being of patients.[21,22]

A result of this realisation was that several broader quality of life measures have been put forward in previous years, intended to be used in health economic evaluations. This includes for example the ICEpop CAPability measure for Older people (ICECAP-O),[23] the ICEpop CAPability measure for Adults (ICECAP-A),[24] the Adult Social Care Outcome Toolkit (ASCOT),[25] and the Well-being Of Older People measure (WOOP).[26] The use of such broader measures in cost-utility analysis implies extending the scope of Q to broader well-being. However, this requires further empirical work and comes with additional challenges, which have not been fully addressed so far. Some of these challenges will be addressed in this dissertation.

Whose values count?

The use of multi-dimensional health or well-being instruments in cost-utility analysis requires deriving weights for all possible health or well-being states, i.e., creating a utility tariff for the instrument. Without weighting, we would assume that each of the included dimensions and underlying levels have the same impact on the calculated health or well-being index.[1] If you take the EQ-5D for example, this would imply that a high level of anxiety and depression is equally bad as a high level in pain/discomfort. This may or may not be an accurate description of the preferences of people for those states. The same goes for the impact of different levels of functioning within one health domain (such as mobility) on quality of life. Utility tariffs are predominantly generated by eliciting preferences towards the different attributes (and their levels) using some form of survey-based experiments. One of the questions that arise in this context, is whose preferences to elicit. In other words, whose values should count when creating a utility index?

Predominantly, preferences were obtained from samples of the general population. The central argument for this choice is that the general population is the payer of health care, and that therefore their preferences should be reflected when valuing health or well-being states to assess the benefits of health care for use in economic evaluations.[6] In the choice experiments that are used to elicit such preferences, individuals are usually asked to imagine being in particular health or well-being states, which they do not necessarily experience themselves (or even have experienced in the past). Therefore, the elicited type of utility is also called ex-ante, anticipated or decision utility. The alternative to this approach is to obtain preferences from individuals who actually experience these health or well-being states, the patients. These are consequently called experienced utility.[27] While different methodologies have been used in the past to elicit experienced utility for health instruments,[28] no such research exists yet for broader well-being measures.

Preference elicitation techniques for creating utility tariffs

While many different techniques for health state valuation exist,[29] the elicitation of decision utility is primarily based on methods grounded in expected utility theory. The main types of methods that have been applied in the past for health utility instruments were standard gamble, time trade-off, or more recently, discrete choice experiments (DCE) and best-worst scaling (BWS).[6] While all of these methods have their own advantages and disadvantages, DCEs have become more widely used for this purpose in recent years.[30,31] The emergence of broader well-being instruments for use in economic evaluations raises the questions whether the same elicitation methods can also be used for such instruments, and how to choose between methods. To approach these questions, the following two features of broader well-being instruments should be highlighted.

A broader scope of Q usually implies that the descriptive systems of the corresponding well-being instruments cover more domains (or dimensions) as compared to traditional health instruments. The ASCOT, WOOP, and the EQ Health and Wellbeing Short version (EQ-HWB-S), which is currently being developed (https://euroqol.org/blog/eq-hwb/), all have nine dimensions.[25,26] The WI-X, another novel well-being measure, which is currently validated, has even 10 dimensions, while ICECAP-A and ICECAP- have five dimension.[23,24] The most commonly used health measures, EQ-5D and SF-6D, have five and six dimensions, respectively.[6] This generally larger set of domains and their levels has two implications: First, the statistical power required to properly estimate utility weights increases considerably. This means that sample sizes need to be larger and/or the efficiency of the elicitation format needs to be increased. Second, the complexity of the experiments themselves increases with every additional dimension or level. Therefore, the cognitive burden of the elicitation format becomes an increasingly important consideration when choosing the valuation method. This is especially relevant if measurement is meant to be inclusive / representative and hence also include members of the population with lower cognitive capacity.

Another aspect worth mentioning here is that of anchoring the utility scale. Anchoring is a necessary feature of the QALY framework, to be able to combine quality of life with length of life.[1] For health instruments and the resulting QALYs, the states dead (0) and full health (1) form the natural anchors (while negative values for states 'worse than dead' are also possible).[32] For broader well-being instruments, this may not be as straightforward. In previous valuation studies of well-being instruments, the question of anchoring was not fully addressed.[25,33] More specifically, placing the state of being dead on a scale from 'no well-being' to 'full well-being' was not unambiguous.

The definition and estimation of v_Q

As outlined above, knowledge about incremental costs and benefits, however defined, is not sufficient to assess whether an intervention is cost-effective when incremental costs and benefits are both positive (quadrant A of Figure 2). Some form of assessment has to be made on whether the additional benefits are worth the additional costs, in other words, provide sufficient value for money. The still acceptable ratio is called the cost-effectiveness threshold, depicted as v_Q in equation 1 and displayed in Figure 2.[1] Some jurisdictions shy away from explicitly formulating a threshold value (in € per QALY).[34,35] However, it is worth noting that any reimbursement decision made includes some implicit consideration of the appropriate ratio of benefits and costs. Arguably, using an explicit threshold, based on some form of societal deliberation and/or empirical evidence, should be the preferred approach, since it increases the accountability and transparency of health care decision making.[36]

Depending on the jurisdiction, the threshold relates to two different concepts. For instance, NICE in England takes a health care perspective in health technology assessment. The corresponding threshold represents the cost-effectiveness of displaced care, or the health opportunity costs. In countries like the Netherlands a broader societal perspective is prescribed by the National Healthcare institute. There, the threshold represents the consumption value of health or, in other words, the societal value of health.[9] The following, and the corresponding chapters of this dissertation, focus on the latter conceptualisation, as this is the relevant one in the Dutch context.

So how can the consumption value of health v_Q be measured? In principle, v_Q and the related threshold should be oriented on preferences of the general population, as the ultimate payer (and consumer) of health care. The elicitation of such preferences can be attempted using two conceptually different approaches. First, one could use people's actual behaviour and choices to obtain an approximation of how much they value their health in monetary terms. Such 'revealed preferences', for example, were obtained based on wage-risk trade-offs (that is, the willingness to accept a higher health risk associated with a job for a wage premium). However, identifying stable and causal estimates is notoriously difficult in this context.[37] Furthermore, such approaches can only estimate how much individuals value their own health. This may not be sufficient to inform collective decision making.

The predominantly applied alternative to such revealed preference estimations is setting up experiments with hypothetical choices and thereby eliciting 'stated preferences'. In the endeavour to estimate v_Q, willingness to pay experiments have been the most widely used approach.[38] However, while flexible in their specification and straightforward to implement, such experiments also have significant disadvantages.[39] This also applies when obtaining estimates of v_Q.[40,41] Therefore, the search for alternative approaches for estimating v_Q is still ongoing and newly developed approaches require further attention.[42–44]

Objectives and structure of this thesis

The first general objective of this thesis is to conduct research relating to broadening the specification of the benefit dimension to well-being, as outlined above, aiming to facilitate the use of broader outcome measures in health economic evaluations. This aim will be addressed in Part I of this thesis, which is entitled **"Generating weights for health and well-being instruments"**. Relating to what has been discussed in the previous section, the second broader objective of this thesis is to apply and refine methodologies aiming to estimate the monetary value of health and broader well-being. This will be the focus of **Part II** of this thesis with the title **"Estimating the monetary value of health and well-being gains".** It is important to note that while the two parts in general answer distinct questions, extending the scope of health economic evaluations to broader quality of life requires adequately answering both sets of questions (and other questions not discussed here). The following outlines the two parts, and the corresponding chapters and research questions.

Part I, "Generating weights for health and well-being instruments", consists of three studies presented in chapters 2 to 4. This part of the thesis discusses and applies alternative methods for generating utility weights for multi-dimensional quality of life instruments, with a special focus on instruments going beyond health.

Chapter 2 focuses on generating utility weights for the ICECAP quality of life instruments. While utility tariffs already exist for the ICECAP, these were based on ex-ante or decision utility, i.e., individuals valuing states that they are currently not in. In this study, we instead created utility values based on experienced utility, i.e., individuals' actual experience of a state. The research question for this chapter was: *Is it feasible to create experienced utility tariffs for the broader ICEACP well-being* *measures with well-being data, and how do these experienced utility tariffs compare to decision utility tariffs?*

Chapter 3 includes a comparison of two types of choice experiments for obtaining utility weights for a broader quality of life measures. The comparison focuses on the cognitive burden of the experiments, which increases with the size of the descriptive systems of the instrument to be valued. Research question: *Is best-worst scaling or discrete choice experiment the more appropriate method in terms of response burden for eliciting utility tariffs for a large well-being instrument?*

Chapter 4 reports on results of a discrete choice experiment eliciting anchored utility weights for a nine-dimensional well-being instrument: the Well-being Of Older People measure (WOOP). The WOOP was designed to be used as a broader outcome measure in economic evaluations of health and social care interventions targeted at older people. Research question: *What are the preferences of older people in the Netherlands regarding the relative importance of the nine well-being dimensions of the WOOP?*

Part II, "Estimating the monetary value of health and well-being gains" is comprised of four studies (chapter 5 to 8). It revolves around estimating *v*, i.e., the monetary value of health or well-being gains.

Chapter 5 reports on a study applying a traditional willingness-to-pay experiment to obtain the societal valuation of health safety provided by an early warning system for infectious diseases. This study was performed before the current COVID-19 pandemic and little did we know back in 2018, when we started this project, that this topic would become this relevant. Research question: *How much are citizens in six European countries willing to pay for an early warning system for infectious diseases?*

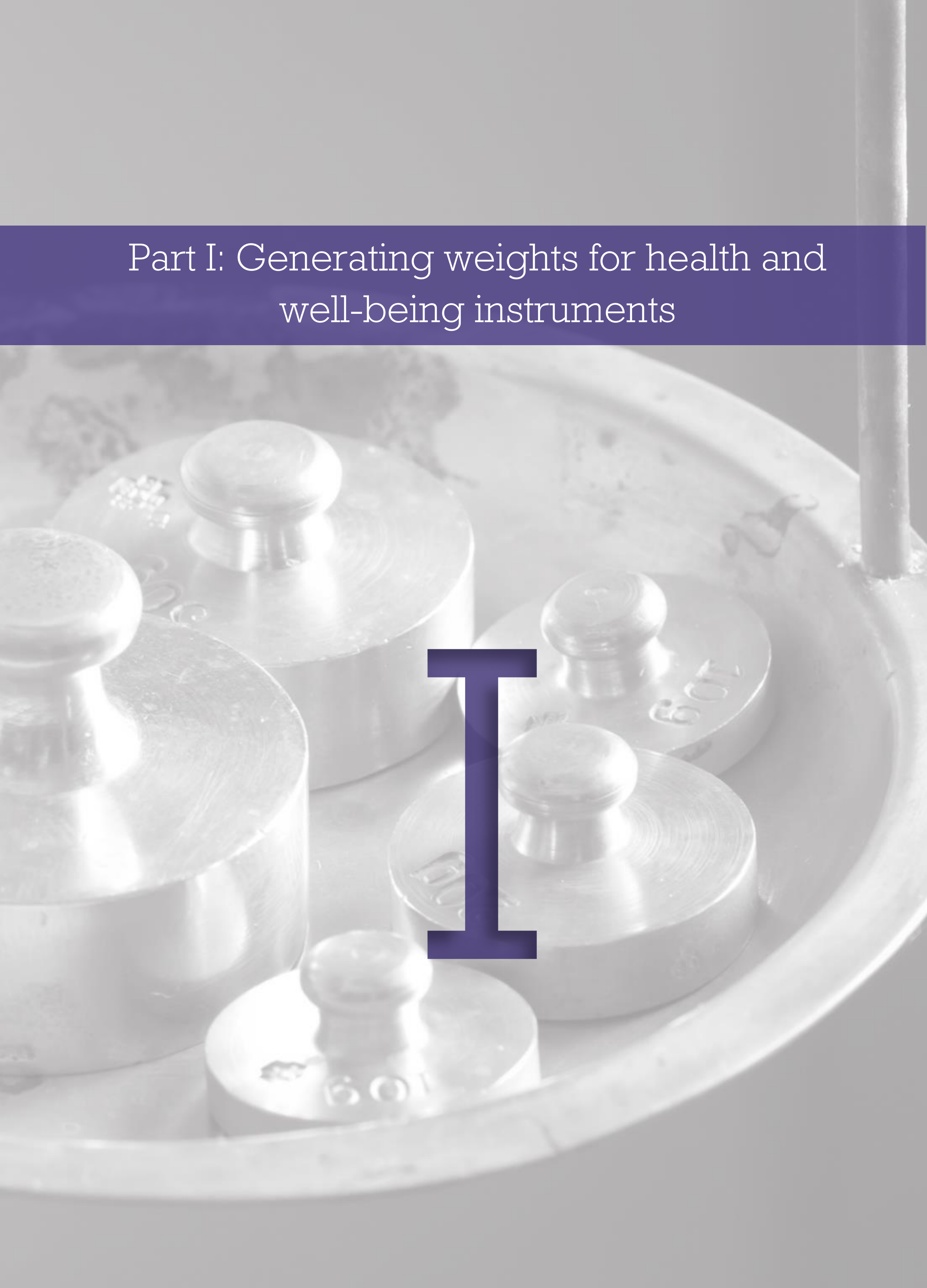
Chapter 6 is a replication of the experiment presented in chapter 5 but fielded in 2020, a few months after the start of the COVID-19 pandemic. Research question: *Did the valuation of an early warning system for infectious diseases change compared to before the pandemic?*

Chapter 7 presents a study which estimated the monetary value of a QALY, but at the same time the monetary equivalent value of broader gains in well-being, more specifically capability well-being. This side-by-side estimation allowed a first direct comparison of the relative value of health and well-being. In this study, we applied the well-being valuation method, an approach only recently suggested for valuing health gains. Research question: *Is it feasible to estimate monetary values for both health and well-being based on this approach, and what is the relative monetary value of health gains and well-being gains?*

Chapter 8 comprehensively tests and refines this regression-based method for valuing health gains. In this study we used large-scale longitudinal data from Germany to empirically assess several open issues when applying the well-being valuation method for estimating the monetary value of health gains. Research question: *What are empirical challenges of the well-being valuation approach for estimating the monetary value of health and how can some of these be addressed?*

To conclude this thesis and bring together the different parts, Chapter 9 summarises and discusses the main results of the research conducted for this thesis. The results are furthermore put into a broader context, highlighting strengths and limitations of the thesis as well as the policy relevance and implications of the work, and avenues for future research.

Part I: Generating weights for health and well-being instruments



I

2

Happy with your capabilities? Valuing ICECAP-O and ICECAP-A states based on experienced utility using subjective well-being data

SFW Himmler, NJA van Exel, WBF Brouwer

Medical decision making, 2020: 40(4):498-510

Abstract

Background: The ICECAP-O and the ICECAP-A are validated capability well-being instruments. For use in economic evaluations, multi-dimensional instruments require weighting of the distinguished well-being states. These weights are usually obtained through ex-ante preference elicitation, i.e., decision utility, but could also be based on experienced utility. Objective. This paper describes the development of value sets for ICECAP-O and ICECAP-A based on experienced utility and compares them to current decision utility weights.

Methods: Data from two cross-sectional samples corresponding to the target groups of ICECAP-O and ICECAP-A was used in two separate analyses. The utility impacts of ICECAP-O and ICECAP-A levels were assessed through regression models using a composite measure of subjective well-being as proxy for experienced utility. The observed utility impacts were rescaled to match the 0 to 1 range of the existing value set.

Results: The calculated experienced utility values were similar to the decision utility weights for some of the ICECAP dimensions but deviated for others. The largest differences were found for weights of the ICECAP-O dimension enjoyment and the ICECAP-A dimensions attachment and autonomy.

Conclusions: The results suggest a different weighting of ICECAP-O and ICECAP-A levels if experienced utility is used instead of decision utility.

Introduction

The allocation of scarce healthcare resources is an important and difficult task for health care decision-makers. In that context, the costs and benefits of competing healthcare interventions are increasingly compared with each other. Typically, such comparisons are supported by health technology assessment, with an important role for economic evaluations.[45] In the healthcare decision making context, the latter often takes the form of a cost-utility analysis in which costs are expressed in monetary terms, while benefits are expressed in terms of quality-adjusted life-years (QALYs). Health-related quality of life is commonly measured by generic multi-dimensional instruments like the EQ-5D. Health states are then valued using utility weights to create an index score anchored at 0 (dead) and 1 (full health).[45,46]

However, it has been questioned whether maximising health, as captured in QALYs, is an appropriate representation of society's values concerning health care,[18] or the appropriate objective in all areas of health care.[24] The benefits of health care in many situations are not limited to health alone. In palliative and elderly care for example, health improvement might not even represent the (primary) aim of interventions.[19,20] Interventions in these areas may be targeted at increasing well-being rather than health. This implies that (part of) the benefits of interventions may not be appropriately captured when using traditional health-related quality of life measures.

The increasing awareness of this issue has led to the development of instruments that allow for a more complete evaluation of health care interventions. Two prominent outcome measures are the ICEpop CAPability measure for Older people (ICECAP-O) and the ICEpop CAPability measure for Adults (ICECAP-A). The ICECAP-O was developed for assessing the capability well-being of older people (65+).[23] The instrument consists of five attributes, namely (i) attachment, (ii) security, (iii) role, (iv) enjoyment, and (v) control. Capability in each domain is measured using four levels. The ICECAP-A instrument aims to measure capability well-being in the general adult population (18+), using five dimensions: (i) stability, (ii) attachment, (iii) autonomy, (iv) achievement, and (v) enjoyment.[24] The validity of ICECAP-O[47–49] and ICECAP-A[50–52] have been studied with generally favourable results, with the caveat that the ICECAP-O may not fully capture physical health.[53]

For use in economic evaluations, multi-dimensional instruments like the ICECAP-O and ICECAP-A do not only require a descriptive system of health or well-being states, but also a valuation or weighting of those states. This weighting allows measured states to be expressed on a 0 (worst well-being state described with the instrument) to 1 (best well-being state described with the instrument) scale. One option to calculate such a set of weights (or tariff) is using general population preferences.[46] The current tariffs for ICECAP-O and ICECAP-A were obtained from representative samples from the respective target populations using best-worst scaling experiments.[33,54] These types of experiments elicit preferences by asking people to imagine being in particular states, which they do not experience themselves. The obtained preference weights, therefore, are based on ex-ante or decision utility.

A much debated question is whether decision utility is the appropriate basis in the context of valuing health or well-being states or whether weights should be derived from people's experience of health and well-being states (experienced utility).[27,55] A key advantage of using experienced utility is that weights need not be based on choices in relation to hypothetical state descriptions, but can be based on the actual experience of the valued health or well-being states. Arguably, this leads to a better understanding of the effect a health or well-being state on overall quality of life.[56] Decision utility and experienced utility can differ substantially, with the valuation of states involving impaired physical health usually being higher when based on experienced utility,[57,58] possibly due to coping and adaptation.[59–61] While both decision utility and experienced utility have their advantages and disadvantages, they may both be relevant for decision-makers.[62]

So far, only tariffs based on decision utility are available for the ICECAP-O and the ICECAP-A *well-being* measures. In the large, but heterogeneous literature regarding experienced utility-based values for *health* states, summarised by Cubi-Molla et al. (2018),[28] different approaches to assess *experienced health* have been proposed, including the visual analogue scale or time-trade-off using the respondent's experienced health state. Although our current study aimed to derive tariffs based on experienced utility for broader well-being states rather than for health states, these different approaches may be relevant in that context as well, especially in deviating from deriving preferences for hypothetical states. For our current study, however, we chose a different, more direct approach to approximate experienced utility, which we deemed to be more appropriate in the context of broader well-being outcome measures. The here applied methodology entailed measuring the correlation of well-being states with subjective well-being (SWB) using regression techniques.[63,64] This approach is derived from the notion that ratings of SWB, or life satisfaction, constitute an informative approximation of the underling and unobservable construct of welfare or utility.[65] In order to capture current *experienced* utility, this type of analysis requires the simultaneous measurement of SWB and the health or well-being instrument. Data provided by the two already provide relevant information on their own on the current experienced well-being state. However, combining that information to obtain an indication of the importance of the instrument's items in terms of experienced utility is arguably producing more pertinent and informative data for measuring the impact of an intervention if one is interested in what dimensions drive the outcome.

The proposed approach has been previously applied to the health state descriptive systems of EQ-5D-3L and SF-6D using general population[66,67] and patient data.[68] Results indicate that differences between decision utility and experienced utility exist. The latter, for instance, gives more weight to mental health compared to pain and physical functioning, arguably because adapting to mental health problems is more difficult.[60]

This paper describes the development of experienced utility tariffs for the ICECAP-O and ICECAP-A instruments based on SWB data from two general population samples from the UK. We compare our results to the existing decision utility tariffs. This information is valuable for the future use of capability well-being in health economic evaluations in contexts where experienced and perceived capabilities are expected to diverge. We furthermore contribute to the discussion of using experienced or decision utility in economic evaluations.

Methods

Data

ICECAP-O and ICECAP-A were developed for the measurement of capability well-being of two different age groups (65+, 18+). Therefore, we used data from two separate cross-sectional surveys. The survey targeted at the elderly was administered to a sample of 516 UK citizens aged 70 and above in 2015 and was initially designed to validate existing well-being outcome measures in the elderly.[49,53] The adult population survey was administered to a sample of UK citizens aged 18 to 65 in 2018. This second sample consisted of 1,373 complete observations. Both surveys were intended to be representative in terms of age, gender, and education, were conducted online, and administered by a sampling agency using quota sampling. The analysis of both instruments followed the same protocol.

Measurement of subjective well-being as proxy for experienced utility

We used SWB data to assess experienced utility. Our datasets contained two widely accepted SWB measures: Cantril's ladder (CL) and the Satisfaction with Life Scale (SWLS). CL is a one-dimensional instrument asking respondents where they would place their life on a ladder ranging from worst possible to best possible life, using a 0 to 10 scale.[69] The SWLS is a five-item measure asking respondents to rate statements like "The conditions of my life are excellent" on a seven-point Likert scale leading to a range of possible values from 5 to 35.[70] While CL has the advantage of being self-anchored and intuitive, the SWLS, due to its multiple items, has higher reliability and facilitates better comparisons across individuals.[71] No clear gold standard has been established for SWB measurement.[71]

Due to the lack of clear guidance and as we did not want to constrain ourselves to one of the two measurements of well-being, we used a composite measure of both instruments, calculated as the unweighted averages of CL and SWLS values, which were rescaled to a 0 to 1 index.

$$\mathrm{SWB_i} = \frac{\mathrm{SWLS_i} + \mathrm{Cantrils\ Ladder_i}}{2} \tag{2}$$

Such a composite measure could arguably be more robust and informative than either on its own.[72] While the two instruments are strongly related, one likely caries SWB information the other measure does not contain.[73] Additionally, combining the results of two instruments measuring the same concept could reduce the impact of response errors. To test the sensitivity of our results to the type of SWB measure selected, we repeated our analyses using CL and SWLS separately.

Statistical analysis

To estimate the relationship between ICECAP-O and ICECAP-A states and SWB, we regressed our composite measure of SWB on all levels of the five ICECAP dimensions for all individuals *i*. Equation (3) contains the model estimated for the ICECAP-O:

$$SWB_i=\beta_0+AT_{il}\beta_{ATl}+SEC_{il}\beta_{SEl}+RO_{il}\beta_{ROl}+EN_{il}\beta_{ENl}+CO_{il}\beta_{COl}+SES_i\beta_{SES}+\varepsilon_i \tag{3}$$

The terms AT, SE, RO, EN and CO represent the vectors containing all dummy-coded levels *l* of the five ICECAP-O dimensions attachment (AT), security (SEC), role (RO), enjoyment (EN), and control (CO), with the highest levels of the dimensions (e.g., "I can have all the love and friendship I want") as reference categories. SES is a vector of variables describing the socioeconomic status of individuals. This vector includes gender, age, education, marital status, financial situation and wealth, which are expected to be related to the SWB of individuals.[74] The model estimated for the ICECAP-A is presented in equation (4):

$$SWB_i=\beta_0+ST_{il}\beta_{STl}+AT_{il}\beta_{ATl}+AU_{il}\beta_{AUl}+AC_{il}\beta_{ACl}+EN_{il}\beta_{ENl}+SES_i\beta_{SES}+\varepsilon_i \tag{4}$$

ST, AT, AU, AC and EN are vectors, which contain the dummy-coded levels *l* of the ICECAP-A dimensions stability (ST), attachment (AT), autonomy (AU), achievement (AC), and enjoyment (EN), with again the highest levels of capabilities as reference categories. SES, the vector of socioeconomic variables, consists of the same variables as in the ICECAP-O model, except for replacing wealth with income, which seems more appropriate in a working-age population sample.

Equations (3) and (4) were estimated using ordinary least squares (OLS), assuming cardinality of the composite SWB values, an assumption that has been shown to hold for the type of SWB measures used in this analysis.[65] To account for the censored nature of the SWB values (0 to 1), Tobit models were also tested. The coefficient estimates were largely similar to the OLS results, but the models were inferior concerning model fit. Functional form specifications of control variables followed model fit. A reduced model only including ICECAP level dummies was estimated to test the robustness of ICECAP-level coefficients to model specification. Given that levels within domains have a natural order, we subjected the model to monotonicity constraints if regression results produced illogical ordering in the level coefficients. In contrast to related studies,[68] a dummy variable indicating the worst level in any dimension was not included in the presented analysis, as the variable was not significant (p=0.571 and p=0.809) and did not influence coefficient estimates in either ICECAP-O or ICECAP-A regressions.

Calculation of tariffs

The coefficient estimates of the full models were used to construct the value sets. As the highest levels of ICECAP dimensions were taken as the reference categories in the OLS regressions, the coefficients of ICECAP-O and ICECAP-A levels represent the disutilities experienced due to being in a particular, lower capability state. The disutilities were linearly rescaled to a 0 to 1 range by summing up the level four coefficients, linearly extending these coefficients to sum up to 1, and multiplying the remaining coefficients with the same factor. Standard errors of the rescaled disutilities were calculated by bootstrap estimation, drawing samples with replacement, and repeating the regression and rescaling steps, setting the number of bootstrap replications to 500. To test whether the disutilities were significantly different to the corresponding values based on decision utility, t-statistics were obtained using the calculated standard errors. The t-tests did not account for the uncertainty in the decision utility weights, as their standard errors were not reported.[33,47] In a final step, the disutilities were reverse coded (e.g. the reference level was changed from 'completely independent' to 'unable to be at all independent') to generate utility values with the utility of 'no capabilities' being defined as 0 (state 44444) and full capability defined as 1 (state 11111). Descriptive analysis, regressions, rescaling, and bootstrapping were performed using STATA 15.0 (Stata Corp. 2018. Stata Statistical Software: Release 15. College Station, TX: Stata Corp LP). Data and STATA code can be made available upon request from the corresponding author. The disclaimed funding source had no role in the study.

Table 1: Life satisfaction, capabilities, and background characteristics of samples.

	ICECAP-O data	ICECAP-A data
Male	53.7%	48.2%
Mean age (SD)	75.1 (4.97)	42.9 (13.7)
Finished tertiary education	45.2%	45.4%
Married	60.1%	59.5%
Make ends meet		
With great difficulty	4.3%	8.0%
With some difficulty	26.2%	37.8%
Fairly easy	42.3%	40.0%
Easily	27.3%	14.2%
Median household wealth	£ 77,500	
Median household income per month		£ 2,250
Mean Cantril's Ladder score (SD)	0.70 (0.19)	0.64 (0.20)
Mean SWLS score (SD)	0.63 (0.22)	0.52 (0.24)
Mean composite SWB score (SD)	0.66 (0.19)	0.58 (0.21)
Mean ICECAP-O/-A score [a] (SD)	0.81 (0.15)	0.75 (0.20)
N	516	1,373

Note: SD, Standard deviation; SWLS, Satisfaction With Life Scale. [a] Using current decision utility value sets.

Results

Descriptive analysis

Table 1 reports on the characteristics of the two samples used for our analysis. The calculated means of CL, SWLS, and the composite SWB measure suggest that the senior population had a higher overall subjective well-being than the sample with people aged 18-65. This result was in line with previous findings.[74] The composite SWB measure naturally averaged out differences between CL and SWLS and had a mean of 0.66 (SD 0.19) and 0.58 (SD 0.21) for the ICECAP-O and ICECAP-A datasets, respectively.

Figure 1 shows the distributions of selected levels per ICECAP-O and ICECAP-A dimension in both data sets. Dimensions with the lowest level of capabilities were security and enjoyment for the ICECAP-O and stability and achievement for the ICECAP-A. In all dimensions, the lowest levels of capability were only selected by between 1.6 and 8.0% of respondents.

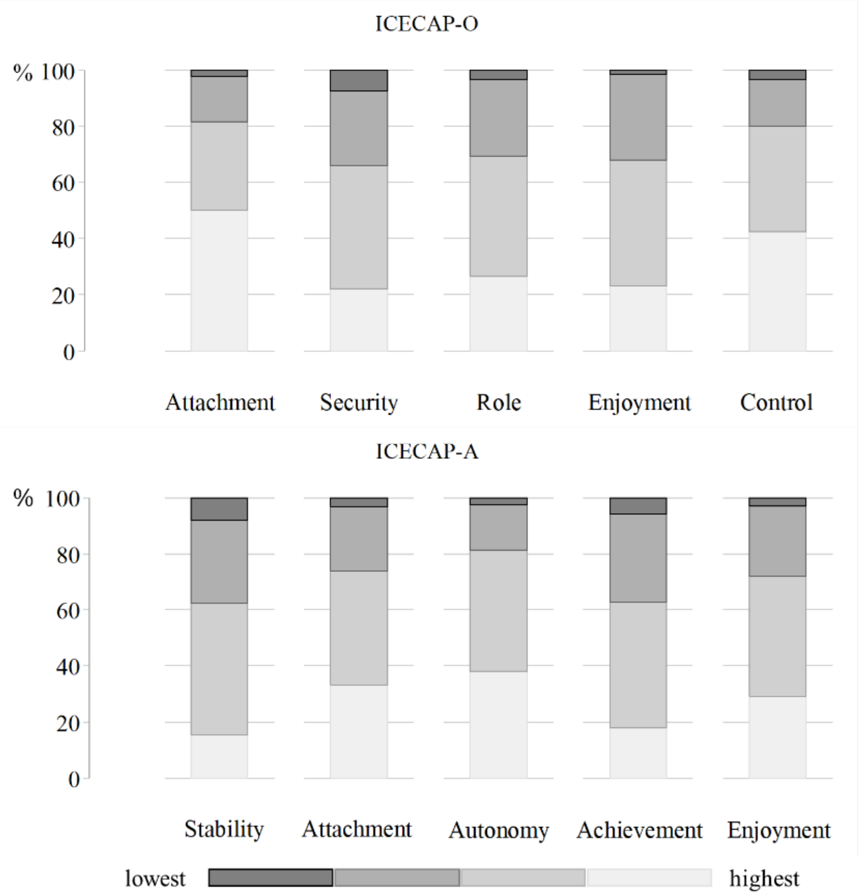


Figure 3: Distribution of selected capability levels per dimension in the two samples.

Results from OLS regressions

The estimation results of the impact of the ICECAP-O and ICECAP-A levels on SWB are presented in Tables 2 and 3. The tables contain both reduced models with only ICECAP levels in column (I) and full models including control variables in column (II). Adding control variables to the two models only slightly changed the size of the ICECAP coefficients, while the improvements in R^2 values from 0.628 to 0.647 and 0.630 to 0.656 were small.

The intercept coefficients for the reduced ICECAP-O and ICECAP-A models did not reach 1 (0.868 and 0.817, respectively), signalling that although capabilities describe a considerable part of SWB, full capability does not imply full SWB. On the lower end of the scale, no capabilities (ICECAP profiles 44444) correspond to SWB of 0.143 and 0.117, respectively. Both instruments, therefore, roughly described 70% of the spread of possible SWB values.

In the full ICECAP-O model (II), shown in Table 2, all ICECAP-O levels were significant at the 5% level, except for the role dimension and the second level of the security dimension. The coefficient of the role 2 variable, i.e., role domain, answering level 2, was positive (0.004), although insignificant. To obtain consistently logical orderings we reran the regression, constraining the role 2 variable to be zero using the STATA command cnsreg. Imposing this constraint only marginally changed the overall coefficients. Being married and having a better financial situation had the expected positive relationship with SWB.[74] For ease of comparison, columns IV and V in Table 2 list the rescaled experienced disutilities of ICECAP-O levels based on SWB data, as well as the decision disutilities from the above-mentioned tariffs, changing the reference category from level 5 to level 1. As column VI shows, the disutilities of the enjoyment levels were larger when calculated based on experienced utility. Further significant differences were found in a lower disutility for level 2 in the security dimension and a higher value for level 3 in the control dimension.

In the full ICECAP-A model (II), shown in Table 3, adding controls changed the coefficients of the capability levels slightly. In this model, the three levels of the autonomy dimension, levels 2 and 3 of the attachment dimension, and level 2 of the achievement dimension were not significant on the 5% level. The attachment levels 2 and 3 were significant in the reduced model (I), but their effect was partly absorbed by adding the controls (coefficients changed from -0.020 to -0.014 and -0.033 to -0.024 for levels 2 and 3, respectively). Being female, married, and having less financial hardship all had the expected significant positive relationship with SWB.[74] Comparison of disutilities based on decision and experienced utility (III & IV) using t-tests revealed sizable and significant differences in all ICECAP-A dimensions except for the achievement dimension (V). Higher experienced disutilities were found for the stability and the enjoyment dimensions and lower experienced disutilities for the attachment and autonomy dimensions compared to the values based on decision utility.

Table 2: Impact of ICECAP-O dimensions on subjective well-being (N= 516)

	(I)		(II)		(III)		(IV)	(V)	(VI)	
	Reduced model		Full model		Constrained		EU [a]	DU [b]	Diff.	p-value[c]
Attachment 2	-0.042**	(0.014)	-0.041**	(0.014)	-0.040**	(0.014)	-0.059	-0.021	-0.038	0.096
Attachment 3	-0.096***	(0.019)	-0.090***	(0.020)	-0.090***	(0.020)	-0.132	-0.120	-0.012	0.723
Attachment 4	-0.188***	(0.045)	-0.164***	(0.046)	-0.164***	(0.046)	-0.241	-0.266	0.025	0.683
Security 2	-0.018	(0.014)	-0.016	(0.014)	-0.016	(0.014)	-0.024	-0.072	0.048**	0.006
Security 3	-0.083***	(0.019)	-0.081***	(0.019)	-0.081***	(0.019)	-0.119	-0.113	-0.006	0.849
Security 4	-0.140***	(0.029)	-0.131***	(0.029)	-0.131***	(0.029)	-0.193	-0.147	-0.046	0.348
Role 2	0.001	(0.016)	0.004	(0.016)	0.000	(.)	0.000	-0.013	0.013	-
Role 3	-0.040	(0.026)	-0.036	(0.025)	-0.039*	(0.019)	-0.057	-0.063	0.006	0.832
Role 4	-0.096	(0.061)	-0.097	(0.062)	-0.099	(0.059)	-0.146	-0.177	0.031	0.712
Enjoyment 2	-0.062***	(0.016)	-0.058***	(0.015)	-0.057***	(0.013)	-0.083	-0.002	-0.081***	<0.001
Enjoyment 3	-0.130***	(0.023)	-0.127***	(0.023)	-0.126***	(0.021)	-0.185	-0.048	-0.137***	0.001
Enjoyment 4	-0.155**	(0.052)	-0.134*	(0.054)	-0.133*	(0.054)	-0.195	-0.149	-0.046	0.553
Control 2	-0.043**	(0.014)	-0.042**	(0.014)	-0.041**	(0.013)	-0.060	-0.025	-0.035	0.085
Control 3	-0.151***	(0.024)	-0.143***	(0.023)	-0.142***	(0.023)	-0.209	-0.102	-0.107*	0.010
Control 4	-0.146***	(0.042)	-0.154***	(0.043)	-0.153***	(0.043)	-0.225	-0.261	0.036	0.604
Male			-0.016	(0.011)	-0.016	(0.011)				
Age			0.010	(0.032)	0.010	(0.032)				
Age-squared			-0.000	(0.000)	-0.000	(0.000)				
Tertiary education			-0.002	(0.011)	-0.002	(0.011)				
Married			0.029*	(0.012)	0.029*	(0.012)				
Make ends meet[d]										
some difficulty			0.062*	(0.030)	0.062*	(0.030)				
fairly easily			0.063*	(0.030)	0.064*	(0.030)				
easily			0.097**	(0.031)	0.097**	(0.031)				
Wealth in million £			0.0002***	(0.000)	0.000***	(0.000)				
Constant	0.868***	(0.011)	0.327	(1.251)	0.321	(1.252)				

[a] Rescaled to 0-1 interval

[b] From Coast et al. (2008) after reversing reference category

[c] Calculated using bootstrapped standard

[d] Reference category: With great difficulty

Note: Standard errors in parentheses; EU, experienced utility; DU, decision utility. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Appendices A and B contain results from regressions including CL and SWLS separately, instead of a combination of the two SWB measures. The composite score levels out the differences between CL and SWLS coefficients, which were highest in the ICECAP-O dimensions security and role and the ICECAP-A domains attachment and autonomy. The differences in regression results were, in general, more prominent for the ICECAP-O calculations. Fewer instances of illogical orderings, fewer insignificant levels, and higher explanatory power of the models were observed when applying the composite SWB measure.

Value sets of ICECAP-O and ICECAP-A

The value sets based on experienced utility are presented in Table 4. The coefficients of the ICECAP levels were used for calculating the tariffs regardless of their level of significance. Analogue to the previously described findings, the largest differences compared to existing decision utility tariffs were found in the ICECAP-O dimension enjoyment and the ICECAP-A dimensions attachment and autonomy. The latter two received considerably smaller weights. Applying these experienced utility tariffs to our data, changed the mean ICECAP-O utility value from 0.814 (SD 0.150) to 0.716 (SD 0.217), and the mean ICECAP-A from 0.748 (SD 0.202) to 0.656 (SD 0.238). The difference in means for the ICECAP-O can primarily be attributed to the considerably lower weights for enjoyment 2 and enjoyment 3 levels, which were selected by around 75% of respondents (Figure 1). The differences for the ICECAP-A partly had their origin in the lower values for level 1 of attachment and autonomy dimension, which represented the most frequently chosen highest capability levels in the data (Figure 1).

Table 3: Impact of ICECAP-A dimensions on subjective well-being (N= 1,373)

	(I)		(II)		(IV)	(V)	(VI)	
	Reduced model		Full model		EU [a]	DU [b]	Diff.	p-value[c]
Stability 2	-0.066***	(0.012)	-0.059***	(0.012)	-0.094	-0.031	-0.063**	0.001
Stability 3	-0.178***	(0.015)	-0.158***	(0.015)	-0.253	-0.121	-0.132***	<0.001
Stability 4	-0.251***	(0.021)	-0.219***	(0.022)	-0.351	-0.223	-0.128***	<0.001
Attachment 2	-0.020*	(0.009)	-0.014	(0.009)	-0.023	-0.039	0.016	0.234
Attachment 3	-0.033**	(0.013)	-0.024	(0.013)	-0.038	-0.131	0.093***	<0.001
Attachment 4	-0.079**	(0.028)	-0.059*	(0.026)	-0.095	-0.252	0.157***	<0.001
Autonomy 2	-0.008	(0.008)	-0.008	(0.007)	-0.013	-0.032	0.019*	0.041
Autonomy 3	-0.014	(0.012)	-0.013	(0.012)	-0.020	-0.105	0.084***	<0.001
Autonomy 4	-0.029	(0.031)	-0.025	(0.030)	-0.041	-0.182	0.141***	<0.001
Achievement 2	-0.021	(0.011)	-0.017	(0.011)	-0.027	-0.022	-0.004	0.781
Achievement 3	-0.080***	(0.014)	-0.071***	(0.014)	-0.113	-0.090	-0.023	0.310
Achievement 4	-0.171***	(0.024)	-0.159***	(0.024)	-0.255	-0.160	-0.095*	0.013
Enjoyment 2	-0.053***	(0.010)	-0.055***	(0.010)	-0.089	-0.027	-0.062***	<0.001
Enjoyment 3	-0.144***	(0.015)	-0.140***	(0.014)	-0.224	-0.112	-0.112***	<0.001
Enjoyment 4	-0.170***	(0.032)	-0.162***	(0.031)	-0.259	-0.184	-0.075	0.160
Male			-0.016*	(0.007)				
Age			0.000	(0.002)				
Age-squared			0.000	(0.000)				
Tertiary education			0.003	(0.007)				
Married			0.030***	(0.008)				
Make ends meet [d]								
some difficulty			0.033*	(0.016)				
fairly easily			0.077***	(0.016)				
Easily			0.094***	(0.019)				
Monthly income £			0.000	(0.000)				
Constant	0.817***	(0.010)	0.710***	(0.041)				

[a] Rescaled to 0-1 interval

[b] From Flynn et al. (2015) after reversing reference category

[c] Calculated using bootstrapped standard errors

[d] Reference category: With great difficulty

Note: Standard errors in parentheses; EU, experienced utility; DU, decision utility. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4: Experienced utility tariffs for ICECAP-O and ICECAP-A.

	ICECAP-O EU tariff	ICECAP-O DU tariff [a]		ICECAP-A EU tariff	ICECAP-A DU tariff [b]
Attachment 1	0.241	0.2535	Stability 1	0.351	0.2221
Attachment 2	0.182	0.2325	Stability 2	0.257	0.1915
Attachment 3	0.109	0.1340	Stability 3	0.098	0.1013
Attachment 4	0.000	-0.0128	Stability 4	0.000	-0.0008
Security 1	0.193	0.1788	Attachment 1	0.095	0.2276
Security 2	0.169	0.1071	Attachment 2	0.072	0.1890
Security 3	0.074	0.0661	Attachment 3	0.056	0.0964
Security 4	0.000	0.0321	Attachment 4	0.000	-0.0239
Role 1	0.146	0.1923	Autonomy 1	0.041	0.1881
Role 2	0.146	0.1793	Autonomy 2	0.028	0.1560
Role 3	0.089	0.1296	Autonomy 3	0.021	0.0836
Role 4	0.000	0.0151	Autonomy 4	0.000	0.0063
Enjoyment 1	0.195	0.1660	Achievement 1	0.255	0.1811
Enjoyment 2	0.112	0.1643	Achievement 2	0.228	0.1588
Enjoyment 3	0.011	0.1185	Achievement 3	0.142	0.0909
Enjoyment 4	0.000	0.0168	Achievement 4	0.000	0.0210
Control 1	0.225	0.2094	Enjoyment 1	0.259	0.1811
Control 2	0.165	0.1848	Enjoyment 2	0.170	0.1540
Control 3	0.016	0.1076	Enjoyment 3	0.035	0.0693
Control 4	0.000	-0.0512	Enjoyment 4	0.000	-0.0026

Note: EU, experienced utility; DU, decision utility. [a] From Coast et al. (2008) [b] From Flynn et al. (2015).

Figure 2a shows the positions and ICECAP index values of four ICECAP profiles on the respective 0-1 scale applying the two value sets. Index scores based on experienced utility are positioned to the left of decision utility scores. The largest differences between the value sets within the four exemplary ICECAP profiles was found for a change from the ICECAP-O state 44444 (no capabilities) to the ICECAP-O state 33333, which increased the utility score from 0 to 0.556 using the decision utility tariffs and from 0 to 0.299 using experienced utility tariffs (i.e., a difference of 0.257). Figure 2b plots ICECAP index values for all observations used in this analysis, with experienced utility values on the x-axis and the decision utility values on the y-axis. These comparisons show that the differences between index values using the two sets of weights are more pronounced for lower utilities and that the discrepancy was larger for the ICECAP-O values.

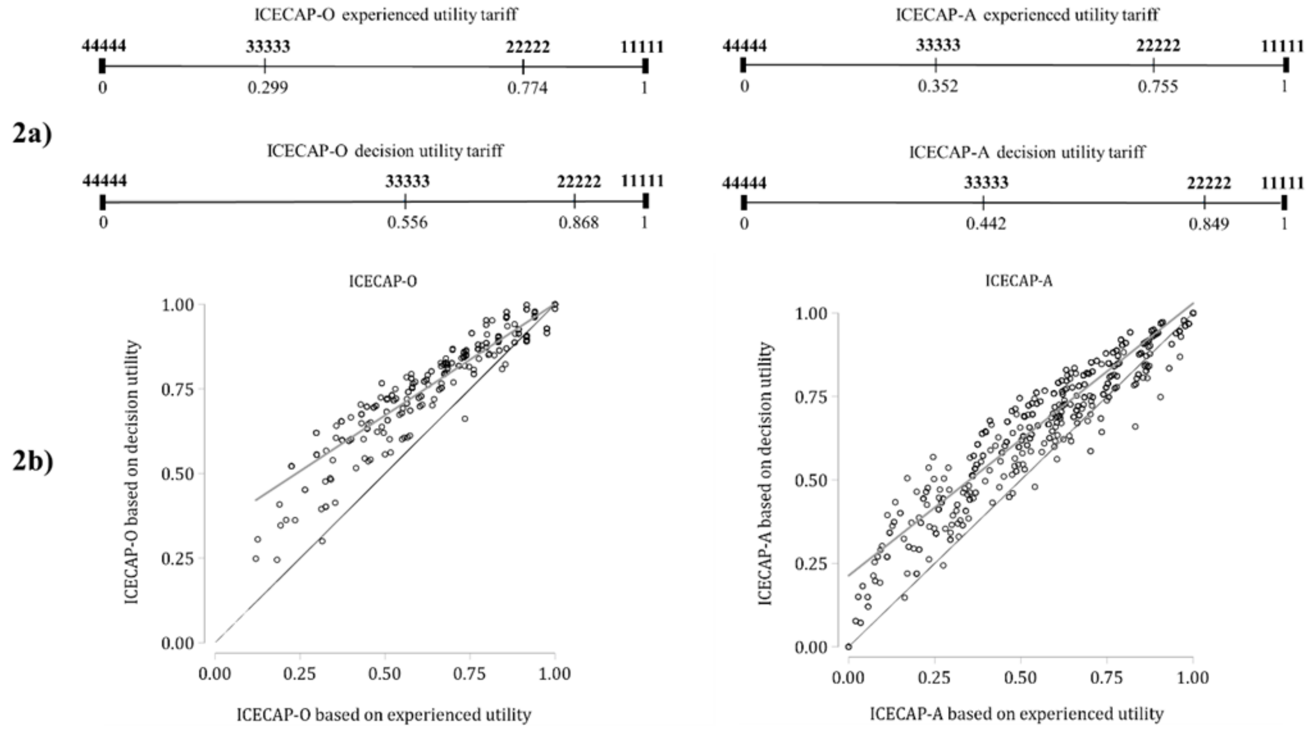


Figure 4: Utility values for main ICECAP profiles and sample population comparing experienced and decision utility tariffs. 44444 indicates no capabilities, 11111 full capabilities. (a) Comparison of position and index values of the main anchors (no capability, full capability) and 2 additional profiles using the 2 value sets. (b) Plot of the ICECAP index values for all observations using the 2 value sets.

Discussion

Summary and context of results

The capability instruments ICECAP-O and ICECAP-A have the potential to broaden the evaluative space of economic evaluations of health care interventions. Levels and dimensions of instruments like the ICECAP-O and ICECAP-A have to be weighted to determine a single utility score that can be used as a measure of benefit in cost-utility analyses.[46] These weights should ideally reflect what matters most to people and can be based on decision utility or experienced utility. This choice is not neutral, as resulting values sets can differ,[67] as they do here. While tariffs based on decision utility are available for ICECAP-O and ICECAP-A, this was not yet the case for experienced utility. Therefore, we developed these, by directly assessing well-being capability values based on their impact on SWB using regression, interpreting life satisfaction as (a proxy of) experienced utility. This is different from approaches often taken in the related literature on self-rated experienced *health*, due to the broader nature of the ICECAP-instruments.[28]

Differences between the existing decision utility tariffs and the here derived tariffs in general are smaller for the ICECAP-O than for the ICECAP-A value sets. A surprising finding was the positive coefficient of the role 2 variable in comparison to role 1, which represented the highest level of capabilities in that domain. However, the coefficient was small (0.004) and not significantly different from zero. This finding may merely indicate little difference between level 2 and level 1 in that specific ICECAP-O dimension in terms of experienced utility. The largest differences in ICECAP-O value sets were found in the enjoyment dimension. In the ICECAP-A value set, the weights of the attachment and autonomy dimensions were considerably smaller than the decision utility weights, while stability and enjoyment dimensions received higher weights.

The observed differences could originate from various aspects. For instance, it could be that respondents performing decision utility exercises overestimate the impact of a specific capability domain on their utility. When occurring in real life, the impact on experienced utility may be smaller, for example, due to easier adaptation.[59,75] Moreover, we observed relatively few people with poor capability in the attachment dimension, which may have reduced statistical power. Finally, loss of autonomy may often occur jointly with other reductions in capability, so that parts of its impact is already captured through other dimensions, which may be more pronounced in experienced utility than in decision utility where respondents need to consider the separate domains. One could also speculate that for some individuals SWB may be negatively related to autonomy as a higher level of independence might be indicative of lacking close relationships or attachment. However, we found no support for this hypothesis in our data, as we observed a positive, significant correlation between autonomy and attachment dimension ($r=0.25$, $p<0.001$).

To our knowledge, our study was the first to analyse the differences between valuations of capability states based on ex-ante decision utility and experienced utility. The existing literature on using the latter to value health states, namely of EQ-5D-3L and

SF-6D, shows that the estimates of the impact of specific dimensions can differ substantially between the two approaches,[67,68,76] especially for mental health problems (e.g. EQ-5D dimensions anxiety and depression). The impact of mental health problems on quality of life is much smaller when based on decision utility than when based on experienced utility. One study, using experienced utility, even estimated the impact of mental health problems to be about ten times larger than the impact of mobility constraints, while these dimensions typically have similar impacts in existing tariffs based on decision utility.[67] In that context, the discrepancies between decision and experienced utility tariffs found in our study were relatively small, in particular for the ICECAP-O.

We also emphasise that the ICECAP-O and ICECAP-A levels explain a considerably larger share of the variation in SWB (R-squared of 0.63 and 0.63) than EQ-5D or SF-6D in a previous analysis (e.g. 0.30 and 0.42 in [68]). The level coefficients furthermore describe a wider spread of possible SWB values than has been reported for EQ-5D and SF-6D in a similar analysis.[68] Both are indications that the ICECAP measures indeed capture broader quality of life than just health-related quality of life.

A novelty in the approach used here is that, instead of using a single one-dimensional life satisfaction score as a proxy for SWB and experienced utility, we constructed a SWB measure based on two well-established measures. When replicating our analysis using the measures separately (see Appendices A and B), we obtained similar coefficients, but the composite measure performed better than the separate measures regarding logical orderings, significant levels, and overall model fit. The use of the composite score appeared to average out differences between SWB measures and may be seen to provide a broader indication of SWB, potentially superior to using the measures separately.

An important issue worth mentioning here, although beyond the scope of the current paper, is that of anchoring the value set. As mentioned, the here presented ICECAP tariffs range from the worst state described by the instrument (i.e., no capabilities, state 44444) to the best state described by the instrument (i.e., full capabilities, state 11111). Hence, the tariffs are not anchored to the state of being dead (sometimes seen as a 'natural zero', in particular for instruments measuring health). This approach is in line with the scoring of the decision utility weights of the ICECAP, which were not anchored on the state of being dead either.[23,47] This makes it unclear how that state of being dead would relate to the scale used here, and marks a clear difference with much of the QALY literature, where anchoring to the state of being dead constitutes a central concept.[46] In general this remains an understudied topic in well-being research and deserves attention in the future.

Limitations

While general limitations and caveats of the chosen analytic approach are discussed elsewhere,[66,77,78] we have to acknowledge the following limitations specific to our study. First, the data used for this study was obtained through online surveys of existing panels. Individuals participating in such online panels might differ from the general population, especially in the elderly. Second, our analysis is based on samples with modest sizes (516 and 1,373), which, for instance, lead to relatively low numbers of observations in the lowest levels of capabilities (Figure 1). Our calculations depend on the discrepancy between coefficients of the lowest and the highest levels. Therefore, our results may be influenced by a limited number of observations regarding particular states.

Furthermore, potential endogeneity issues in our models also deserve emphasis. A reverse causal relationship between health and SWB has been shown to exist.[79] It is not unlikely that this also the case for capability well-being, although future research using longitudinal or experimental data needs to confirm this. Our results could further be biased by omitting variables relevant to SWB. Our dataset did not include variables capturing personality traits, social environment, or community involvement. All of these can be important predictors of SWB and likely to be correlated with the level of capability well-being, or their perception by individuals.[74]

Lastly, the approach we applied here, is based on preferences and utility, which may, to some extent, conceptually be considered to be at odds with adopting the capability approach. Amartya Sen, who developed the capability approach, explicitly rejected the (exclusive) focus on emotional responses to states to determine their value, for instance arguing that preferences adapt to circumstances and are prone to psychological biases and effects.[80] Nevertheless, the previously established ICECAP tariffs were also based on preferences as at present there seems to be no feasible and superior alternative approaches in valuing capability well-being states.[33,47]

Implications

The here estimated utility weights share a fair degree of similarity with the decision utility weights, and more so for the better capability states than for the worse states (see Fig 2b). This creates some confidence that the chosen approach produces relevant valuations and deserves further attention. Nonetheless, aggregating the weights into specific states can produce significant differences between the two value sets (see Figure 2a). We do not know to what extent these differences result from the different measurement approaches or the different concepts that were measured (experienced versus decision utility). This also implies that it is unknown how the here obtained estimates relate to 'true' (unobserved) underlying experienced utilities - which is also true for the existing decision utility value set. Future research could investigate and disentangle these issues further.

We reported the differences in the mean ICECAP scores applying the different tariffs (see Results section) and in contrast to previous analyses for health found that the utility levels of individuals are lower when applying experienced rather than decision utility tariffs. Future studies could investigate this interesting result further, preferably in larger datasets and among patients, as one possible explanation for the current finding is that it may be driven by relatively few observations of very poor ICECAP states (see Figure 1). Furthermore, these differences and their implications should be interpreted with caution as they represent different constructs. Experienced utility incorporates coping and adaptation to well-being states,[61] which decision utility probably does not. The presented tariffs, which were rescaled on a 0 to 1 range, tell us something about the relative weight of different levels in different dimensions, which as such can be compared to the relative weights from the decision utility value set. However, given that both experienced and decision utility relate to different underlying constructs, it would be inaccurate to claim that a similar absolute change based on the two tariffs indeed have the same underlying unobserved utility impact. This is because the utility scales underlying decision and experienced utility are not the same (e.g., due to adaptation).

Notwithstanding this, the comparison of value sets does highlight that a choice for either tariff set can have important consequences for evaluations. Applying the tariffs based on experienced utility would entail putting more weight on some ICECAP dimensions and less on others when assessing the benefits of an intervention, as compared to using tariffs based on decision utility. Moreover, the tariffs based on experienced utility appear to result in a more even spread of the capability states on the scale. The findings shown in Figure 2a imply that the decision utility tariffs give much weight to moving people from the worst capability state (44444) to the state with poor capabilities in all domains (33333), i.e., a gain of 0.556 and 0.442 respectively for the ICECAP-O and ICECAP-A. The same improvement would be assigned a utility gain of 0.299 and 0.352, respectively, if the experienced utility tariffs were applied. These differences might have considerable implications for the assessment of interventions achieving such a change. Similarly, an improvement from state 22222 to the best capability state 11111 receives more weight when using the tariffs based on experienced utility as compared to those based on decision utility: 0.226 versus 0.132 for ICECAP-O, and 0.245 versus 0.151 for ICECAP-A. Such differences highlight the importance of an informed choice on which tariffs to use to inform allocation decisions.

As this is to a large extent a normative choice, we advocate applying the here presented experienced utility-based tariffs alongside the decision utility-based tariffs for the UK context, as knowledge about the actual SWB impacts of experiencing certain states can be useful complementary information for decision making.[62] We do advocate more research to confirm the validity of the here derived sets in that context. In general, the application of ICECAP measures as substitutes for or complements of health-related quality of life measures in different contexts requires further research.

Furthermore, we recommend broader use of SWB valuation approaches and presenting experienced utility as well as decision utility impacts of interventions where available and relevant. Moreover, in cases where obtaining a value set based on decision utilities is (too) difficult or costly, the here used approach may be a reasonable and relatively straightforward alternative to produce relevant valuations of health or well-being states.

Conclusions

Concluding, our analysis showed that calculating value sets for the ICECAP-O and ICECAP-A instruments based on experienced utility using SWB data is feasible and that the obtained weights to some extent differ from the weights previously obtained based on decision utility. This difference generates insights for policymakers in the context of the application of ICECAP-O and ICECAP-A as well as experienced and decision utility in economic evaluations.

3

What works better for preference elicitation among older people? Cognitive burden of discrete choice experiment and case 2 best-worst scaling in an online setting

SFW Himmler, V Soekhai, NJA van Exel, WBF Brouwer

Journal of Choice Modelling. 2021; 38 (100256)

Abstract[*]

To appropriately weight dimensions of quality-of-life instruments for health economic evaluations, population and patient preferences need to be elicited. Two commonly used elicitation methods for this purpose are discrete choice experiments (DCE) and case 2 best-worst scaling (BWS). These methods differ in terms of their cognitive burden, which is especially relevant when eliciting preferences among older people. Using a randomised experiment with respondents from an online panel, this paper examines the cognitive burden associated with colour-coded and level overlapped DCE, colour-coded BWS, and 'standard' BWS choice tasks in a complex health state valuation setting. Our sample included 469 individuals aged 65 and above. Based on both revealed and stated cognitive burden, we found that the DCE tasks were less cognitively burdensome than case 2 BWS. Colour coding case 2 BWS cannot be recommended as its effect on cognitive burden was less clear and the colour coding led to undesired choice heuristics. Our results have implications for future health state valuations of complex quality of life instruments and at least serve as an example of assessing cognitive burden associated with different types of choice experiments.

[*] The online version of this article (https://doi.org/10.1016/j.jocm.2020.100265) includes an appendix with materials relating to appendices B and C in this chapter.

Introduction

Developments like ageing populations and rapid advances in medical technology create challenges for budgets of publicly funded health care systems.[2] Policy makers increasingly have to decide about which health care services to include in the basic benefits package, which should only be made available to certain subpopulations, and which should not be funded at all. Health technology assessment (HTA) generates valuable insights to support this decision-making process, using tools like cost-utility analysis. There, the benefits of health technologies are typically expressed in the amount of (health-related) utilities they produce. These utilities are calculated based on data from generic, multidimensional quality of life instruments, and a weighting algorithm for the levels of the dimensions based on population or patient preferences. The use if these instruments allows calculating health benefits in terms of Quality-Adjusted Life-Years (QALYs), a common metric in cost-utility analyses combining length and quality of life.[46] Given that health and social care, for instance aimed at older persons, may affect more than health-related quality of life alone, more recently, broader well-being measure have been developed.[81] These facilitate cost-utility analyses with a broader scope in terms of relevant outcomes but require obtaining preferences for different 'well-being states'.

The measurement of population and patient preferences in health care is a rapidly developing field, with a plethora of qualitative and quantitative methods to the disposal of researchers and practitioners.[29] One of the most popular methods over the last decade was the discrete choice experiment (DCE). Increasingly, population and patient preferences in health care are obtained using DCEs.[31] The 'standard' DCE entails asking respondents to choose between two or more alternatives[82] and is widely used for weighting quality of life instruments (Mulhern et al., 2018).[30] Generating such weights is also called 'health state valuation'. Another preference elicitation approach that gained traction over the last years also in this context, is best-worst scaling (BWS). There are three different forms of BWS – object case, profile case, and multi-profile case. The following will focus on profile case, or also called case 2 BWS, where individuals have to select a best and a worst option from a list of dimension levels or items.[83] Case 2 BWS was applied to value different quality of life instruments before.[84] This includes the ICECAP-O, a well-being measure specifically aimed at older people.[33]

While both DCE and BWS provide numerical estimates of the relative importance of the different levels and dimensions of the respective quality of life or well-being instrument, previous research directly comparing DCE and BWS has shown that the choice between these approaches is not neutral as resulting preference estimates can differ.[85] According to a recent review comparing DCE and BWS, there seems to be no conclusive evidence yet on which of the methods should be preferred in terms of the validity of the estimates.[86] Both methods assume different choice processes and ultimately may be seen to answer more or less subtly different questions. Some researchers prefer DCEs because the modelled choice processes have a strong theoretical foundation in random utility theory.[87] Providing choices between multiple alternative profiles can also be considered as a more realistic way of the decision-making process compared to selecting a best and worst option from a list of items. On the other hand, some argue that profile case BWS is to be preferred as it is a more efficient way of collecting data compared to DCE since each task entails two choices. Moreover, cognitive burden of BWS tasks may be lower, since individuals only need to focus on one set of attributes and levels in each choice task, compared to multiple in DCEs. Some specifically claim that it would be recommendable to choose case 2 BWS if DCE tasks are considered to be too burdensome.[88,89] However, Whitty and Oliveira Gonçalves (2018) conclude that there is no clear evidence for an advantage of BWS regarding participant acceptability in terms of feasibility of administration or response efficiency. The response efficiency, that is, the cognitive burden associated with choice tasks, is important as it influences choice consistency, respondent fatigue and the use of simplifying choice heuristics,[90] which could subsequently influence the validity of the preference estimates.

Due to the ageing of the population, the need for economic evaluations of health and social care services targeted at older people can be expected to increase. This makes accurately measuring and weighting quality of life dimensions in this population very important, and choosing the appropriate methodology to do so, all the more relevant. Additionally, since there is a large variation in the level of cognitive abilities within older people, the design of choice experiments for this population should especially be wary of the complexity and subsequent cognitive burden of the choice task format in order to enable obtaining valid and reliable responses.[91] Measuring and weighting quality of life or well-being outcomes inaccurately may ultimately lead to sub-optimal policy recommendations for resource allocation to health or social care services aimed at older people.

Specific evidence about the cognitive burden of DCE and case 2 BWS in the context of valuing quality of life measures among older people is lacking. Therefore, the main aim of this study was to assess the cognitive burden and incidence of simplifying choice heuristics in DCE and case 2 BWS choice tasks among older people in this context. Another aim was to test the impact of the use of colour coding on the cognitive burden and choice behaviour of case 2 BWS tasks, which has been assessed for DCEs before.[90]

Methods

We set up a randomised experiment with three study arms to examine the cognitive burden and choice behaviour attached to three respective choice task formats for valuing a quality-of-life instrument: a colour coded DCE, a case 2 BWS and a colour coded case 2 BWS. In the remainder of the paper, BWS refers to case 2 BWS. The quality of life measure used in the experiment was the recently developed Well-being of Older People instrument (WOOP).[92] Examining the cognitive burden of a valuation task is especially important in the context of this new instrument for measuring the general/overall quality of life of older people: First, the WOOP consists of nine dimensions with five levels each, which requires complex choice tasks. Second, as preferences should be based on an older population, cognitive burden is of special relevance. The profiles shown to respondents in both DCE and BWS tasks corresponded to well-being states, described using the nine dimensions of the WOOP (i.e., physical health, mental health, social life, receive support, acceptance and resilience, feeling useful, independence, making ends meet, living situation). Chapter 4 contains an updated version of the full descriptive system of the WOOP, with some formulations differing slightly compared to the version used in this study.

In designing the choice tasks and their visual representation, we followed methodological work on the use of colour coding and level overlap in DCEs aimed to reduce task complexity.[90,93,94] To enable a more direct comparison and to test the impact of colour coding on task complexity in BWS, which has not been studied before, the randomised experiment included a colour coded BWS and a regular BWS. Important to note here is that the design was generated to test the cognitive burden and choice behaviour of older people, not to provide model estimates for the different methods. Due to the large descriptive system of the WOOP, this would have required estimation of 36 parameters in the DCE and 45 parameters in the BWS, and a much larger sample size. While a comparison of model estimates would have been interesting, this was not our current research aim.

Survey structure and randomization

The structure of the experimental survey is shown in Figure 1. First, respondents were asked to complete the WOOP instrument to become familiar with its dimensions and levels. Afterwards, they were randomized 1:1:1 to the three study arms: colour coded DCE (1), colour coded BWS or BWSc (2), and regular BWS (3). The randomization was preferred over having the same respondents completing both DCE and BWS tasks, to have avoid the different parts of the experiment influencing each other and to stay as close as possible to standard DCE and BWS experiments. Furthermore, two full sets of valuation tasks per respondents were considered to be too burdensome. Respondents were familiarized with the presentation of well-being states in the subsequent experiment by showing them their own profile in DCE or BWS format based on the answers they previously gave to the WOOP instrument. The choice task formats were introduced by a simple DCE or BWS task, where participants had to select between two types of fruits or chose the best and worst type of fruit from a list. The second part of the warm-up comprised of a choice task, as used in the subsequent experiment, providing further instructions. Subsequently, a block of six choice tasks was administered, followed by two simple break questions on an unrelated topic to interrupt the monotony and reduce respondent fatigue of answering the choice tasks. Then, a second block containing seven tasks concluded the randomized part of the questionnaire, leading to a total of 13 choice tasks per respondent. All respondents subsequently had to fill in three blocks of evaluation questions on a 5-point Likert scale, before providing some sociodemographic information at the end of the survey.

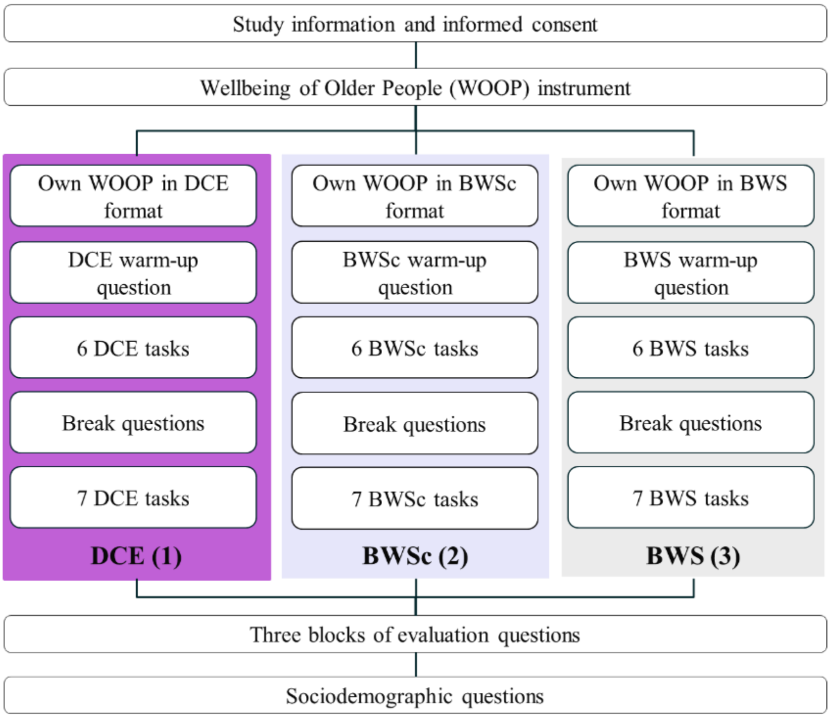


Figure 5: Survey sturcutre and experimental arms

Survey administration and participants

The survey was programmed using Sawtooth software version 9.7.2 (Sequim, WA). We used Prolific.co to recruit survey participants, a platform for online subject recruitment specifically for research purposes.[95] Given our aim to assess the cognitive burden of the choice tasks in a sample of older people, being aged 65 or above was used as inclusion criteria (which is also the target population of the WOOP). Since this age group was underrepresented in the online panel, we had to combine respondents from the two largest country panels of Prolific.co, UK and U.S. residents, to obtain a reasonably sized sample. At the time of data collection, in October 2019, the potential respondent pool contained around 1,000 individuals. Using quota sampling, we aimed for 150 respondents for each of the three study arms. Respondents received a monetary compensation for participating, which was oriented on the mean completion time and averaged to an aggregated hourly reward of £7.62. To test the functionality of the survey and whether respondents understood the choice tasks, six think-aloud interviews with UK residents aged 65 and above were conducted (two per study arm) prior to the main data collection. These interviews showed that participants understood and appropriately engaged in the choice tasks (i.e., traded-off or considered multiple items).

Experimental design of DCE and BWS

Attributes and levels in the DCE and items of the BWS were based on the dimensions and levels of the WOOP instrument (see Chapter 4). This created a rather complex DCE setup with nine dimensions with five levels each and a BWS instrument with 45 items. WOOP well-being states were consequently defined by selecting one of the five levels from each of the nine dimensions for both DCE and BWS. In the DCE, respondents were repeatedly presented with two well-being states and asked to indicate, which of the two they preferred. An opt-out option was not included as this is uncommon in DCEs for health state valuation [30]. In the BWS, a list of nine well-being items corresponding to one well-being state was shown to respondents. Participants then had to select the aspect that they most preferred (best) and the aspect that they least preferred (worst). 'Most' and 'least' is one of the options that are used for describing a best and worst choice.[96] 'Most' and 'least' may have a slightly different interpretation than 'best' and 'worst', but this should not have an impact on cognitive burden.

To ensure that the choice tasks had a similar level of complexity compared to a regular choice experiment, choice tasks were created using standard design methodology as outlined in the subsequent paragraph. The literature on health related DCEs specifically targeted at older people was reviewed (in total 22 papers were studied) to inform the number of choice tasks. The number of choice tasks per respondent varied between 6 and 16 with a mean of 9.2. We opted to select a number of choice tasks at the upper end of this range (13) to capture fatigue effects (examples of this literature are e.g. Arendts et al., 2017; Franco et al., 2016; Milte et al., 2014)[91,97,98] and because we anticipated this might be close to the approximate number in the actual valuation study of the WOOP.

Ten DCE choice tasks were created applying a Bayesian efficient design programmed in Ngene design software (Version 1.2.1) and optimized for a conditional logit, main effects model (online Appendix C contains the utility function) with 36 parameters corresponding to four of the five levels of each of the nine dimensions of the WOOP instrument. Small priors ranging from 0 to –0.25 were assumed, following the logical ordering of the WOOP levels. Besides the think-aloud interviews no further pilot testing was conducted. Level overlap was imposed in five of the nine dimensions to reduce task complexity, as has been shown by Maddala et al. (2003) and Jonker et al. (2018).[93,94] For that purpose, Ngene required a dataset including all possible candidate sets, i.e. combinations of two health states with five overlapped levels. To pragmatically reduce this to a feasible number, 5,000 out of the 1,953,125 possible health states were randomly selected and combined in MATLAB (MathWorks). Out of the obtained 25 million possible sets, we excluded the ones without the specified amount of overlap and randomly selected 1,000 sets out of the remaining 386,030 overlapped sets. Ngene was used to select 10 choice tasks out of the 1,000 candidate sets.

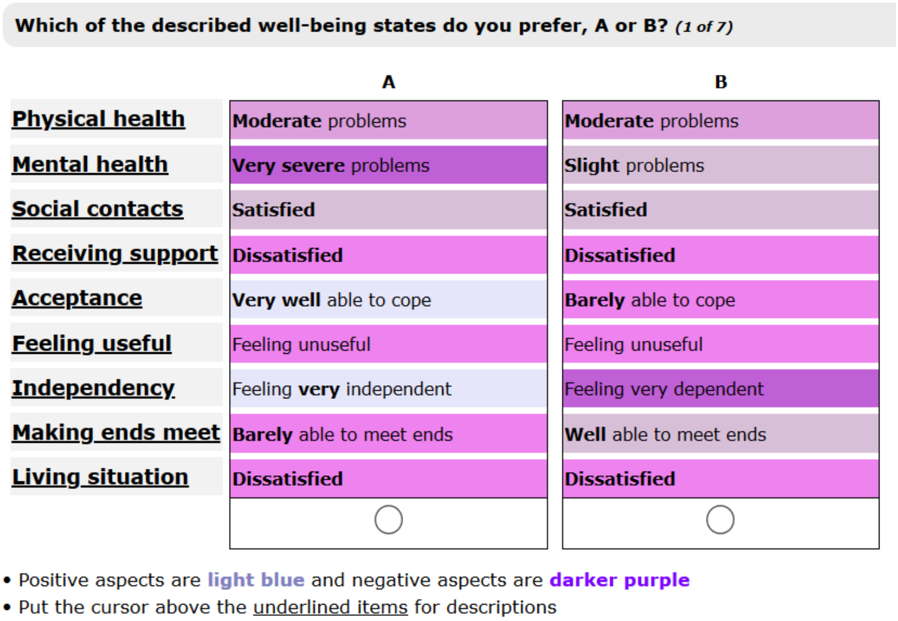
An orthogonal main effects plan using Sawtooth software version 9.7.2 (Sequim, WA) was applied to generate 1,000 blocks of 10 choice tasks for the BWS experiment. Multiple levels from the same WOOP dimension were prohibited to appear in the same task. Following Flynn et al. (2015),[54] to prevent uninformative sets, we reduced the occurrences of tasks with either only one top or bottom WOOP level by deleting all versions where this occurred more than 3 times in the 10 tasks. Out of the remaining 78 versions, one version was randomly selected to be used in the experiment.

We selected one of the created DCE and BWS choice tasks to appear as the second choice task and repeated the tasks at position 8 and 13, to test choice consistency, adding two choice tasks to the original 10 created tasks. In order to reduce the amount of noise in the answers, we chose tasks, which were expected to have a certain degree of utility difference between profiles in the DCE arm or provided somewhat clear BWS choices (the repeated choice tasks are shown in online Appendix B). When this task was repeated the second time, the colour coding of the BWS task was intentionally designed to mislead respondents to assess the dependence on the colour scheme. A dominant DCE choice task and a BWS task, which was expected to have a clear best and worst choice were additionally created and added at position 6 to test the attention level of respondents, adding a third and final choice task to the original ten created tasks. We decided against including results of this task in the final analysis, as such tasks are inherently difficult to compare between DCE and BWS.[86] The order of the dimensions (or attributes) was the same for all respondents and fixed for both DCE and BWS tasks to further reduce task complexity. All respondents received the same 13 DCE tasks in study arm 1. Respondents in study arms 2 and 3 received the same 13 BWS tasks.

Visual presentation of choice tasks

The general visual representation of the choice tasks followed current practice, with the exception that colour coding was added to the choice tasks in study arms 1 and 2. Different shades of purple represented the different attribute levels, with the darker shades of purple highlighting the worse and the lighter shades and light blue expressing the better WOOP attribute levels in both the DCE and the colour coded BWS tasks. This type of colour coding was previously used for DCEs by Jonker et al.,[90,93,99] and was found to reduce task complexity as well as attribute non-attendance and was especially effective in combination with attribute level overlap. The purple colour scheme was specifically designed to accommodate for the most prevalent forms of colour blindness. Additionally, shades of purple do not prompt natural or perceived value judgements, as opposed to for example traffic light colour coding.

Figure 2 shows an example of the layout of the colour-coded (light blue to deep purple) and overlapped (five out of the nine dimensions) DCE choice task. Level descriptions of the WOOP instrument (Chapter 4) were shortened for clarity, level labels were highlighted in bold, and attribute descriptions appeared merely as mouseovers on the attribute labels to reduce the amount of text. Figure 3 shows examples of both colour coded and non-colour coded BWS tasks. Descriptions of attributes were also included as mouseovers, while the item text contained the full WOOP level descriptions.

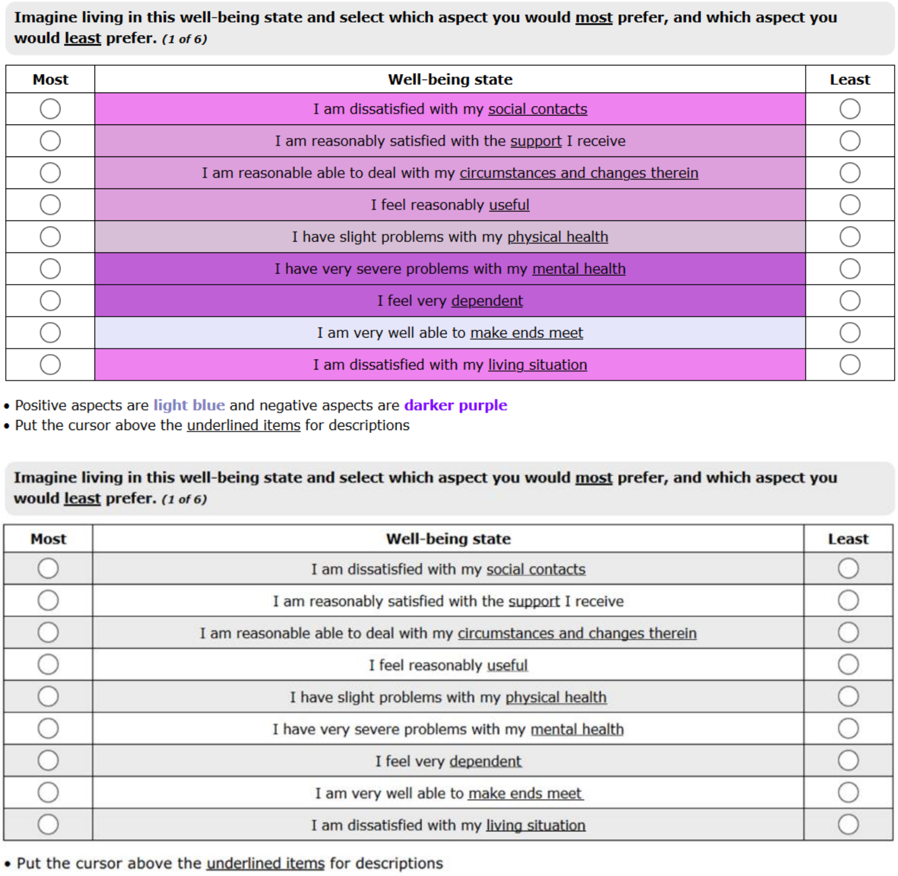


Which of the described well-being states do you prefer, A or B? ***(1 of 7)***

	A	**B**
Physical health	**Moderate** problems	**Moderate** problems
Mental health	**Very severe** problems	**Slight** problems
Social contacts	**Satisfied**	**Satisfied**
Receiving support	**Dissatisfied**	**Dissatisfied**
Acceptance	**Very well** able to cope	**Barely** able to cope
Feeling useful	Feeling unuseful	Feeling unuseful
Independency	Feeling **very** independent	Feeling very dependent
Making ends meet	**Barely** able to meet ends	**Well** able to meet ends
Living situation	**Dissatisfied**	**Dissatisfied**
	○	○

- Positive aspects are **light blue** and negative aspects are **darker purple**
- Put the cursor above the underlined items for descriptions

Figure 6: Visual presentation of DCE choice task with colour coding and level overlap.



Imagine living in this well-being state and select which aspect you would most prefer, and which aspect you would least prefer. *(1 of 6)*

Most	Well-being state	Least
○	I am dissatisfied with my social contacts	○
○	I am reasonably satisfied with the support I receive	○
○	I am reasonable able to deal with my circumstances and changes therein	○
○	I feel reasonably useful	○
○	I have slight problems with my physical health	○
○	I have very severe problems with my mental health	○
○	I feel very dependent	○
○	I am very well able to make ends meet	○
○	I am dissatisfied with my living situation	○

- Positive aspects are **light blue** and negative aspects are **darker purple**
- Put the cursor above the underlined items for descriptions

Imagine living in this well-being state and select which aspect you would most prefer, and which aspect you would least prefer. *(1 of 6)*

Most	Well-being state	Least
○	I am dissatisfied with my social contacts	○
○	I am reasonably satisfied with the support I receive	○
○	I am reasonable able to deal with my circumstances and changes therein	○
○	I feel reasonably useful	○
○	I have slight problems with my physical health	○
○	I have very severe problems with my mental health	○
○	I feel very dependent	○
○	I am very well able to make ends meet	○
○	I am dissatisfied with my living situation	○

- Put the cursor above the underlined items for descriptions

Figure 7: Visual presentation of colour-coded and plain BWS choice task.

Statistical analysis

To assess and compare the cognitive burden and possible choice heuristics associated with the three formats of choice tasks, three types of data were analysed. First, objective measures including mean choice task completion time, development of time per task (assessing learning effects) and drop-out rates were calculated and compared. Second, mean response scores of the three blocks of debriefing questions on perceived choice complexity, the number of choice tasks, and choice strategies used, were obtained. Questions relating to choice strategies were designed to obtain information on the number of attributes commonly considered during the choice tasks, relating to the simplifying heuristic know as attribute non-attendance,[100] and the level of trading between choice tasks.

Third, revealed cognitive burden regarding choice consistency and (simplifying) choice behaviour was assessed based on the actual choices of respondents. This included calculating the proportion of respondents providing the same answers to the twice repeated choice task. For the BWS arm, a consistent response was defined as providing the same answer for either best or worst option, following Krucien et al. (2017).[85] Furthermore, we estimated a lexicographic score, which provides information on trading between attribute levels and dominant choice behaviour. This score was obtained also following an approach applied by Krucien et al. (2017): First, the proportion of choices based on one attribute on an individual level was calculated. Assuming respondents exhibit dominant preferences for an attribute given proportions above 90% (DCE) and 50% (BWS), the lexicographic score was obtained by calculating the proportion of respondents with such preferences.

To test the impact of colour coding on the choice behaviour and strategies in the BWS study arms, the shares of responses based on top and bottom levels of the WOOP dimensions were calculated. Additionally, results from the second repeated choice task, where the colour coding was intentionally misleading, was used to assess the dependence on the colour scheme.

Statistical significance was assessed using Wilcoxon-rank sum tests for the Likert scale data (based on recommendations from de Winter and Dodou, 2010)[101] and chi-squared tests or Fisher exact tests for proportions. A significance level of 10% was used throughout the analysis. Stata 15 (StataCorp 2017) was used for all calculations.

Results

Sample characteristics, dropouts, and completion time

A total of 477 participants successfully started with the experiment and were randomly allocated to the three study arms. No respondent dropped out in study arm 1 (DCE). One of the three dropouts in study arm 2 (BWSc) occurred during the choice tasks and two afterwards. Of the five respondents dropping out in study arm 3 (BWS), four dropouts occurred during answering the BWS tasks and one at a later stage. Fisher exact tests indicated that the difference in total drop-out rates was significantly lower in study arm 1 compared to study arm 3 (0% vs. 3.2%, p-value = 0.029). The difference to study arm 2 was not significant (0% vs. 1.9%, p-value = 0.248).

The characteristics of the remaining sample, split by study arm, are shown in Table 5. The randomisation led to well-balanced samples regarding most sociodemographic aspects, health status (EQ-5D-5L) and well-being (WOOP). 63.7% of the overall sample was younger than 70 years, 34.6% was aged between 70 and 79 years, and 1.7% were aged 80 years and above with 87 years as the maximum age observed.

Table 5: Main characteristics of analysis sample per study arm.

	DCE	BWSc	BWS
	(1)	(2)	(3)
Age in years	69.3	69.1	68.9
Female (%)	0.65	0.60	0.62
Years of education	16.1	15.8	15.8
Country of residence: UK (ref. U.S.) (%)	0.57	0.54	0.52
Employed (%)	0.33	0.29	0.28
EQ-5D-5L utilities (0-1)	0.83	0.82	0.82
WOOP (Sum score rescaled to 0-1)	0.81	0.79	0.82
Number of completes (N)	159	158	152

Note: EQ-5D-5L tariff from Devlin et al. (2018).[32]

The average time it took respondents to complete all 13 choice tasks was 6.0 minutes (SD 3.1) for the DCE tasks, 7.6 minutes for the colour coded BWS tasks (SD 4.9) and 7.2 minutes for the standard BWS tasks (SD 4.6). T-tests indicated that choice task completion time was significantly lower for the DCE tasks compared to the two sets of BWS tasks (p<0.001 and p= 0.007). Figure 4 plots the mean completion time for each choice task separated for each study arm. Differences were most pronounced in the beginning with choice task completion time following a downward trend, likely resulting from learning effects. Respondents in study arm 1 on average answered each choice task faster compared to the BWS study arms, except for one choice task. Differences within the two BWS study arms were less pronounced with the notable exception of choice task 13, where the colour coding was intentionally misleading.

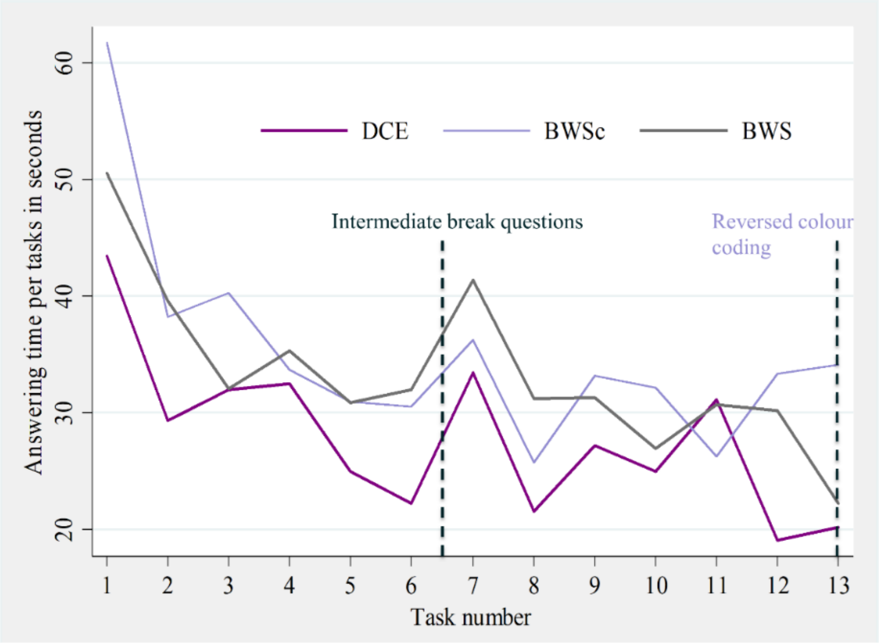


Figure 8: Completion time per choice task within each study arm.

Self-reported cognitive burden of tasks and number of choice tasks

Mean response scores of the three blocks of debriefing questions and results from significance tests comparing the mean scores across study arms are shown in Table 6. DCE choice tasks appeared to be superior in terms of clarity of the tasks and whether tasks were comprehensible from the beginning. Respondents found the presented states easier to image in the BWS tasks, which admittedly confronted participants only with one well-being state instead of two in the DCE. Colour coded BWS choice tasks were evaluated to be less clear than non-colour coded BWS tasks.

Results from the second block of questions indicated that participants from the DCE study arm found the number of choice tasks easier to manage, were more able to stay concentrated over all choice tasks, and could have answered more tasks, compared to the BWS study arms, with most differences being statistically significant. Colour coding the BWS tasks appeared to have a positive effect on the number of choice tasks participants could handle.

Table 6: Mean response score of cognitive debriefing questions.

Question on Likert scale	DCE	BWSc	BWS
from 1 to 5 (5=strongly agree)	(1)	(2)	(3)
Self-reported cognitive burden			
The choice tasks were clear	4.45$^{\dagger\text{ALL}}$	4.11$^{\dagger\text{ALL}}$	4.25$^{\dagger\text{ALL}}$
I could easily choose between the alternatives	3.55	3.65	3.62
I fully understood the choice tasks from the beginning	4.75$^{\dagger\text{ALL}}$	4.26$^{\dagger 1}$	4.36$^{\dagger 1}$
The tasks got easier after answering several	3.77	3.87	3.84
I found some of the presented states difficult to imagine	3.43$^{\dagger 3}$	2.97$^{\dagger 1}$	2.84$^{\dagger 1}$
Number of choice tasks			
The number of choice tasks was manageable	4.64$^{\dagger 3}$	4.54	4.50$^{\dagger 1}$
It was difficult to stay concentrated over all choice tasks	1.72$^{\dagger 3}$	1.94	1.92$^{\dagger 1}$
I could have answered more choice tasks	4.07$^{\dagger\text{ALL}}$	3.91$^{\dagger\text{ALL}}$	3.66$^{\dagger\text{ALL}}$
Answering another block of six 6 choice tasks would be manageable	4.43$^{\dagger\text{ALL}}$	4.19$^{\dagger 1}$	4.18$^{\dagger 1}$
Choice strategies			
I compared all dimensions/items before making my choice	4.72	4.77	4.79
I decided all dimensions/items are equally important	2.86$^{\dagger 3}$	3.00	3.20$^{\dagger 1}$
I always used the same 1 or 2 well-being dimensions to make my choice	3.04$^{\dagger\text{ALL}}$	2.65$^{\dagger 1}$	2.57$^{\dagger 1}$

Note: † $p < 0.10$ of Wilcoxon rank-sum test comparing study arms 1, 2, and 3.

Choice strategies and choice behavior

Most respondents strongly agreed with the statement that they compared all dimensions/items before making their choices, with no significant differences between study arms (Table 6). There were mixed results concerning the perceived level of trading and attributes attended comparing DCE and BWS study arms. While DCE participants agreed to a lesser extent that all dimensions/items are equally important, an indication of trading behaviour, they also reported to a larger degree to having based their decisions on the same 1 or 2 well-being dimensions, which implies some level of attribute non-attendance.

Table 7 lists results for the analysis of choice behaviour. The lexicographic score (see section 2.5), was significantly lower in DCE respondents, indicating more trading and less dominant choice behaviour. In the DCE, dominant preferences were observed only for the physical health attribute. In the BWS, such behaviour was also observed for the mental health and making ends meet attributes, with physical health still being the most prevalent one.

In the DCE study arm, 4.4% of respondents did not provide the same answer to the repeated choice task, when it appeared again for the first time (position 2 and 8), with the same colour code. When it was repeated again as the last choice task, that share was 2.5%. Up to 20% of respondents did not provide either the same best or worst answer in the repeated BWS tasks. It has to be acknowledged, though that the likelihood of providing the same answer by chance alone is larger for DCE choice tasks (50%). When defining consistency as providing the same answer to both best and worst, this share increased to around 60%. There were no significant differences between BWS study arms regarding the choice consistency of the first repeated instance. Almost half of respondents did not provide a consistent best or worst answer to the repeated BWS choice task, where the colour coding was intentionally misleading (position 13). This share was 72.8% when defining consistency in terms of selecting the same best and worst items.

The average individual share of best and worst answer based on either the top and bottom levels of the WOOP dimensions was between 60 and 75%, with higher values observed for the colour coded BWS tasks (significant difference for ‘best’).

Table 7: Revealed choice behaviour

	DCE (1)	BWSc (2)	BWS (3)
Non-trading/dominant choice behaviour			
Lexicographic score	28.9%[†ALL]	79.1%	80.1%
Choice consistency			
% failed a consistent response to repeated choice task (1st) [a]	4.4%[†ALL]	19.6%[†1]	17.8%[†1]
% failed a consistent response to repeated choice task (2nd) [a]	2.5%[†3]	46.8% [b]	19.1%[†1]
% who did not provide same answer for best and worst (1st)		58.9%	61.2%
% who did not provide same answer for best and worst (2nd)		72.8%[§]	60.5%
Focus on top and bottom levels			
Mean individual % of choosing level 1 as best		70.5%[†3]	59.9%[†2]
Mean individual % of choosing level 5 as worst		76.3%	69.4%

Note: [†] $p < 0.10$ of chi-squared tests comparing study arms 1, 2, and 3 (if applicable). [a] For BWS defined as providing either the same best or worst answer. [b] Choice task with intentionally misleading colour coding.

Discussion

To assess the cognitive burden of different types of choice tasks for valuing well-being states for quality-of-life measures in older people, a randomised experiment was conducted, allocating respondents to either a DCE, a colour coded BWS, or a regular BWS format. Our study contributes to the literature by providing empirical evidence on 1) whether DCE or BWS choice tasks are associated with lower cognitive burden in the context of health or well-being state valuation in an older population sample, and 2) whether colour coding of BWS tasks affects cognitive burden and to a lesser extent validity of BWS experiments.

Finding a lower drop-out rate and lower choice task completion time in the DCE study arm compared to the BWS study arms implies that, for older people, DCE choice tasks are less tiring and faster to complete than BWS tasks. Lower completion time was also observed by van Dijk et al. (2016). In terms of self-reported measures, our results indicate that the DCE tasks also were perceived as less cognitively burdensome, and that a higher number of DCE choice tasks was regarded as more acceptable than was a higher number of BWS tasks. The former has also been reported in related studies in different contexts.[86] The latter is especially relevant to consider when thinking about the number of choices per respondent, and hence the required sample size, when selecting DCE or BWS format. Finding lower cognitive burden associated with DCE tasks compared to BWS tasks, in general, is at odds with what has been reported before by Netten et al. (2012).[102] They also compared cognitive burden of DCE and BWS tasks for valuing a large descriptive system of a quality-of-life instrument, but the design of their study was fairly different. The authors used cognitive interviewing, a qualitative approach, in a small sample (N=30), split the DCE task into two parts to reduce the difficulty of the task and showed both DCE and BWS tasks to respondents. Although it does not become clear from the paper, whether respondents had to answer full sets of choice tasks or only one task per method. Whether the difference in findings relates to the differences in design of the studies, is difficult to say.

In terms of (simplifying) choice strategies and choice behaviour, which co-occur with larger cognitive burden, our results are mixed regarding the self-reported behaviour, and less clear cut. We did observe a considerably higher choice consistency and lower degrees of dominant choice behaviour for DCE respondents, with their measurement to some degree accommodating for the methodological differences. However, these results may relate more to artefacts of the type of choice task and may be unrelated to cognitive burden. As stated also by Whitty et al. (2014),[103] the probability of answering consistently to a DCE task by pure chance is already 50%. With nine dimensions this probability is much lower (22%) for the BWS task (defined as providing either the same best or worst answer). Nevertheless, finding that around 60% of BWS respondents did not provide the same best and worst answers when a choice was repeated for the first and the second time, is somewhat worrisome on its own. A higher degree of trading and lower degrees of dominant choice behaviour in DCEs were also reported in the related literature before with a similar caveat as for analysing choice consistency.[85,103]

Comparing colour coded with non-colour coded BWS, we found a similar drop-out rate for both tasks (1.9% and 3.2%, respectively). In the study by Jonker et al. (2018) (study arms 1 and 2), colour coding of the DCE tasks decreased the dropout rate from 13.9% to 9.8%.[93] Further results from the same study set up showed that colour coding alone did not lead to differences with respect to the self-reported cognitive debriefing questions.[90] Our results for BWS regarding these questions are mixed. While participants of the colour coded BWS on average agreed to a higher extent that they could have answered more choice tasks, the non-colour coded BWS choice tasks appeared to have been clearer to respondents. Given no conclusive evidence on cognitive burden, and the fact that the colour coding increased the already high focus on top and bottom levels of the quality-of-life instrument in the BWS tasks, colour coding BWS cannot be recommended for health or well-being state valuation studies among older people.

The overall implications of our analysis must be interpreted considering the following limitations: First, the rather small sample size did not provide us with enough statistical power to be able to use several blocks of choice tasks, which then also would have allowed us to estimate DCE and BWS models. During the design stage, we aimed for 150 respondents per study arm due to the small overall pool of individuals aged 65 on online platforms. While the choice sets were created according to standard design methodology, it could be the case that either of the two choice sets is more difficult to answer in general, irrespective of choice task format, due to smaller utility difference within the shown profiles. As utility weights for the WOOP are not available yet, it was not possible to account for that in the selection of choice set. This risk could have been reduced if multiple blocks would have been used. Related to this, as DCE and BWS models were not estimated, it was not possible to examine differences in utility weights and their potential link to cognitive burden.

Second, per online platform rule, the recruitment of respondents involved a monetary compensation which is rather high compared to standard online panels and which can be reduced if the researcher is not satisfied with the quality of responses. While this is a good thing for respondents and their motivation, this led to very low dropout rates and could have also affected other parts of the analysis. This may reduce the generalisability of our results to studies in this population using other online panels, which by now represent the main source of participants for such experiments. A third caveat of our analysis is that the applicability of our results to the comparison of DCEs without overlap and colour coding, and BWS is limited. However, the use of level overlap in similar DCEs as strategy to reduce task complexity seems to be increasing.[104,105]

Conclusions

Overall, we found evidence that level overlapped and colour coded DCE choice tasks are less cognitively burdensome than BWS choice tasks, in a complex health (or, here, well-being) state valuation exercise among older people. This has implications for future valuation studies, especially since the complexity of the measures to be valued seems to increase when moving from health-related to overall quality of life; see, for instance, the WOOP (Chapter 4), the current plans of the E-QALY project (https://scharr.dept.shef.ac.uk/e-qaly/), or another ongoing study to develop a quality of life measure for older people.[106] Cognitive burden should be an important factor in deciding about which method to choose for valuing such descriptive systems, but at the same time, statistical and theoretical aspects need to be considered as well. Although our results may not be easily generalisable to other topics of study within or outside health care and to other study populations, our study may at least serve as a good example of how to assess cognitive burden associated with different types of choice experiments.

4

Estimating an anchored utility tariff for the well-being of older people measure (WOOP) for the Netherlands

SFW Himmler, MF Jonker, F van Krugten, NJA van Exel, WBF Brouwer

Social Science & Medicine. 2022: 301, 114901.

Abstract*

Objective: Health economic evaluations using common health-related quality of life measures may fall short in adequately incorporating all relevant benefits of health and social care interventions targeted at older people. The Well-being of Older People measure (WOOP) is a broader well-being measure that comprises nine well-being domains. The objective of this study was to estimate a utility tariff for the WOOP, to facilitate its application in cost-utility analyses.

Methods: A discrete choice experiment (DCE) with duration approach was set up and fielded among 2,012 individuals from the Netherlands aged 65 years and above. Matched pairwise choice tasks, color-coding and level overlap were used to reduce the cognitive burden of the DCE. The choice tasks were created using a Bayesian heterogeneous D-efficient design. The estimation procedure accommodated for nonlinear time preferences via an exponential discounting function.

Results: The estimation results showed that 'physical health' and 'mental health', and 'making ends meet' were the most important well-being domains for older people, followed by 'independence' and 'living situation'. Of somewhat lesser importance were domains like 'social life', 'receiving support' and 'feeling useful'. The generated utility tariffs can be used to translate well-being states described with the WOOP to a utility score between -0.616 to 1.

Conclusions: This study established a tariff for the WOOP, which will facilitate its use in economic evaluations of health and social care interventions targeted at older people, first of all in the Netherlands.

* The online version of this article (https://doi.org/10.1016/j.socscimed.2022.114901) contains supplementary materials relating to appendix tables A1 to A3 and figures A1 to A2 in this chapter. The data and code for the analysis presented in this chapter can be accessed through the Open Science Framework: https://osf.io/ysajr/.

Introduction

Health care, social care and long-term care spending is increasing worldwide,[107] propelled by the interaction of ageing populations, increased public expectations, and advances in medical technology.[2] In high income countries, health care spending in the age group above 65 years is already two to three times higher compared to spending in all other age groups combined.[108] Therefore, the efficient use of scarce care resources, especially within this age group, is crucial. Health economic evaluations, like cost-utility analyses, are established tools to assess whether care services are offering value for money and, therefore, are worthwhile investing in. The results of such analyses guide policy makers in their endeavour to provide the best possible care from the available budget. So far, cost-utility analyses predominantly use quality-adjusted life-years (QALYs) as outcome measure, which combine health-related quality of life (HRQoL) with length of life.[46]

Especially in long-term care, social care and end-of-life care, which often aim to improve (or preserve) quality of life domains beyond health, generic HRQoL measures may fall short of measuring the full benefits of these services.[81] As a result, different well-being measures have been developed that aim to capture these quality of life domains beyond health.[81,109–111] However, in developing these measures, lay perspectives on what is important for the well-being of older people have often been overlooked,[112] as well as the heterogeneity in older people's views on what constitutes well-being.[92] Moreover, some of the existing well-being measures are very lengthy and, therefore, not well-suited for self-completion. Most also lack a utility tariff to reflect the relative importance of their domains to overall well-being.[5] While measures like Adult Social Care Outcomes Toolkit (ASCOT) and the ICEpop CAPability measure for Older people (ICECAP-O) do not seem to have these shortcomings,[23,102] questions remain about their evaluative scope.[81] For instance, these measures do not directly measure the quality of life domain 'health'[23,102], even though older people consider this to be (very) important for their well-being.[92,113] While health supposedly is captured indirectly in the ICECAP-O, research suggests that this may not be sufficiently the case, in particular physical health.[26,114,115]

To overcome some of the shortcomings of existing well-being measures, an alternative measure was developed: the Well-being of Older People measure (WOOP).[116] Its domains are directly based on the views of older people in the Netherlands themselves on what constitutes well-being[117] and covers a comprehensive set of nine well-being domains: 'physical health', 'mental health', 'social life', 'receiving support', 'acceptance and resilience', 'feeling useful', 'independence', 'making ends meet', and 'living situation'. For each of the domains, respondents can indicate their level of functioning by selecting one of five response categories (see Appendix). Qualitative research confirmed the content validity and feasibility of the WOOP as it demonstrated that it captured the important domains of well-being for older people and was considered clear and suitable to self-report their level of well-being.[116] Quantitative research showed satisfactory to good results for construct, convergent and discriminant validity, as well as test-retest reliability.[26]

Utility tariffs for the WOOP are currently lacking, which clearly hampers its application in (economic) evaluations of health and social care services for older people.[46] Therefore, the objective of this study is to estimate a Dutch WOOP utility tariff. The structure of this paper is as follows: the next paragraph specifies the methods, with an emphasis on the design of the choice experiment and the data collection; subsequently, the results are presented, including the WOOP utility tariff; finally, we discuss our findings and their implications.

Methods

A discrete choice experiment was designed to estimate utility tariffs for the WOOP for the Netherlands. More specifically, a 'DCE with duration' approach was employed, entailing including duration of life as an additional attribute in the choice tasks This allows anchoring of utilities on a scale from 0 (dead) to 1 (perfect well-being).[118] This method was preferred over standard gamble and time-trade-off approaches due to concerns relating to the cognitive burden of these iterative procedures, the size of the WOOP instrument, and due to the possibility of administering DCE tasks online.[119]
The traditional estimation approach for DCE with duration data assumes linear time preferences. This implies that the general public is willing to give up a constant proportion of remaining life years for a certain health improvement, without consideration of the number of life years that remain.[120] Previous work provided evidence that this assumption does not hold in DCE with duration data and that it would introduce biased parameters, as health state preferences would be contaminated by time preferences Furthermore, we applied a previously developed methodology accounting for non-linear time preferences, which have been shown to exist in DCE with duration data, to avoid biased parameters.[121,122] As such, we did not want to presume linear time preferences from the outset and selected an approach that can accommodate non-linear time preferences with a more flexible approach.[122] How this was achieved is outlined below under 'conceptual framework'.

Attributes, levels, and matched choice task

Attributes and levels used in the choice experiment were defined by the descriptive system of the WOOP (see Appendix).[26] Each of the nine domains of the WOOP is represented by one item with five response levels, generally ranging from excellent (level 1) to bad (level 5). Physical health level 1, for instance, represents being very satisfied with one's physical health. In addition to the nine WOOP domains, a duration attribute was included to enable trade-offs between quality and duration of life. Duration was specified in years using 17 values (0.25, 0.5, 1, 2, …, 15). The values and the range thereof were selected to provide realistic quantities of remaining life years in our target population (smallest and highest values were designed to appear less frequently than the more commonly occurring and hence more realistic middle values). To further increase realism in the choice tasks, we ruled out that the following attribute levels could appear together: Level 1 of independence together with either level 5 of physical or mental health, as well as level 1 of social life and level 5 of support. In a previous data collection with 1,113 respondents, the first two combinations did not occur in the data, while the latter occurred just once.[26]

To reduce the cognitive burden of the ten-attribute choice task for the target population, we undertook several steps. First, descriptions of domains and levels were carefully simplified by the researchers involved in the development and qualitative work of the WOOP instrument. Full domain descriptions were still accessible to respondents in the choice task upon moving the cursor over the abbreviated versions. Second, a previously used matched pairs choice task format, which was found to reduce the cognitive burden of choice tasks, was applied (Figure 9).[99,122] This entailed a first choice between two well-being states A and B, both with equal duration, followed by a matched second choice between the same well-being state B and perfect well-being. This format already simplified the choice tasks by avoiding simultaneous comparisons between the quantity and quality of well-being. This feature of this choice task format additionally helps respondents to treat health and duration multiplicatively. This is theoretically required, but not the case for most respondents when using a traditional, single choice, DCE with duration format.[123] To further reduce the complexity of the choice tasks, five out of the nine domains were constrained to be overlapped (i.e., well-being states differed in only four domains). To highlight the differences, the level descriptions were colour-coded using shades of purple (with darker shades representing worse levels). This combination of level overlap and colour-coding successfully reduced drop-out rates and attribute non-attendance in earlier studies.[90,93] The second choice was between the same well-being state B and perfect well-being, but with a shorter duration.

We confirmed the feasibility of the final choice tasks in think-aloud interviews among individuals aged 65 years and above. In the executed think-aloud protocol, users were asked to verbalize their thoughts as they completed the full concept online survey, in which the DCE was embedded (for the elements of the survey see "Data collection and survey design"). Data saturation was reached after four think-aloud interviews. Obtained information was summarised into three meaningful categories: instructions choice tasks, instructions other tasks, overall layout. Based on the corresponding insights, minor changes were made to the layout of the survey and to the instructions accompanying the warm-up choice tasks.

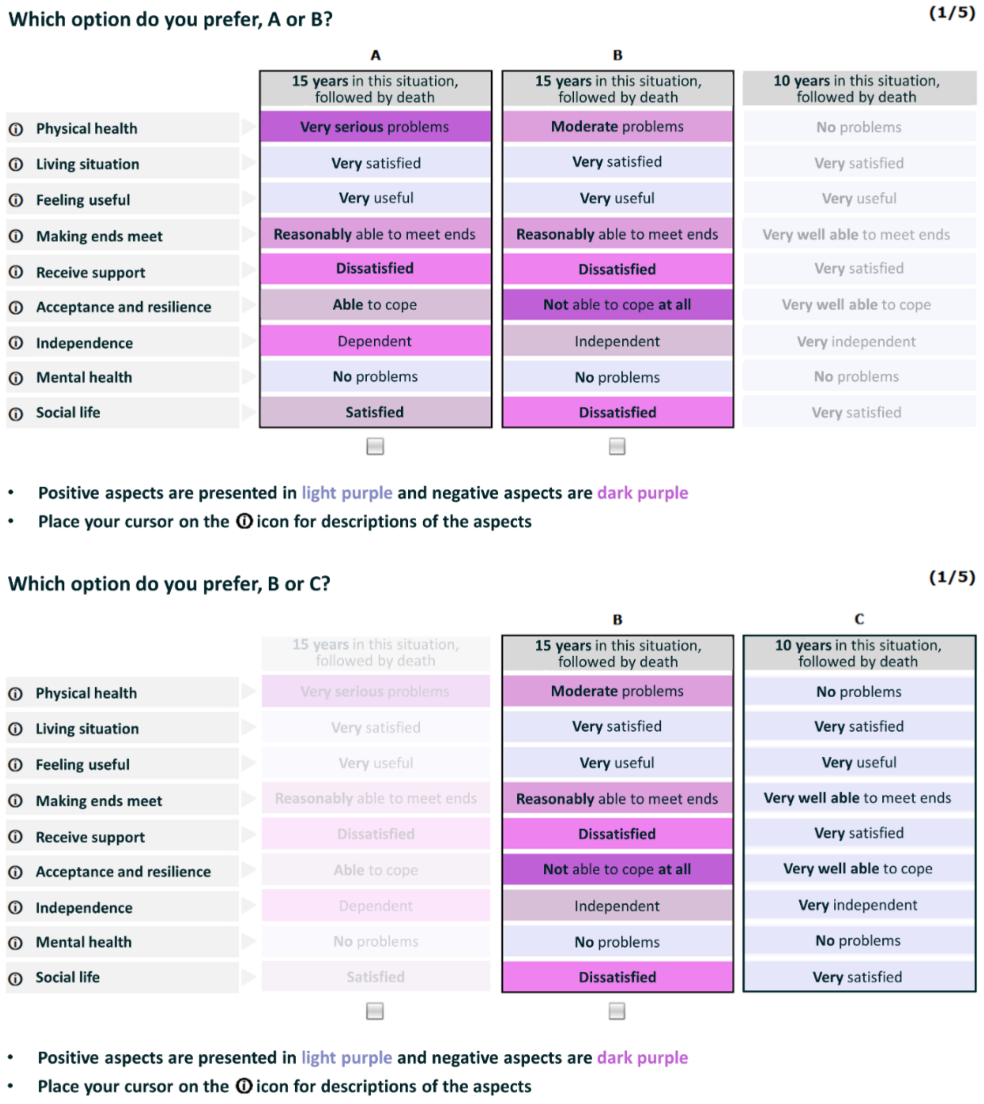
Experimental design

Optimizing the statistical efficiency of the DCE design was crucial due to the large descriptive system of the WOOP (a total of 1,953,125 possible well-being states), and the imposed level overlap constraints. Therefore, an efficient design was implemented and optimized using the TPC-QALY software package.[124] More specifically, a Bayesian heterogeneous D-efficient design with ten sub-designs was used. This implied a simultaneous optimisation of the efficiency of ten separate designs, as well as the efficiency of their aggregate. To give more detail, the D-efficiency criterion was calculated with 100 Bayesian draws, assuming an exponential discount function, and based on the weighted average of the overall (i.e., combined) D-error (0.25) and D-errors of the individual blocks (0.75). An exponential discount function was assumed, which appeared to be the most efficient discount function tested with the TPC-QALY software.[124] The design was optimized for the above-described matched choice task format (see also Figure 1). The number of matched choice tasks per respondent was set at 15, resulting in ten versions and a total set of 300 paired comparisons between two well-being states. Priors for optimising the initial design were informed by logit model estimates WOOP best-worst scaling data (N = 310) from a previous study (Himmler et al., 2020a). The experimental design was updated twice after calculating priors based on 201 and a total of 514 completes to further increase the efficiency of the design.

Data collection and survey design

The DCE was embedded in an online questionnaire and administered to citizens in the Netherlands aged 65 years and above recruited from the panel of the market research company Dynata. We aimed to sample around 2,000 respondents, representative in terms of age and gender, using stratified sampling. After completion, respondents could make a small donation to a charity of their choice. Data collection took place between December 2020 and March 2021.

The survey started with a description of its purpose and a consent form. Next, respondents had to rate their well-being using the WOOP. The DCE training procedure started with a two-alternative choice task with three (randomly selected) WOOP domains. Subsequently, the complexity of the introductory choice task was increased step by step. First, colour-coding was introduced. Second, the duration attribute was added. Third, alternative C and, therefore, the second of the pairwise choice tasks was included. Fourth, all nine WOOP domains were included. Colour-coding, level overlap, and duration were explicitly described. Respondents were randomised to one of the ten blocks of 15 choice tasks between two well-being states. To avoid ordering biases, further randomisation took place regarding the order of choice tasks, the order of the well-being states within choice tasks (A and B), and the order of WOOP domains across respondents (constant order per respondent). The 15 choice tasks were split in three blocks of five tasks, interrupted by two sets of standard socio-demographic questions to reduce response fatigue with respect to the choice tasks. The questionnaire ended with cognitive debriefing questions, an inquiry into whether COVID-19 changed the importance of the WOOP domains for respondents' well-being, and measures for health (EQ-VAS) and life satisfaction (Cantril's ladder).



Which option do you prefer, A or B? (1/5)

	A	B	
	15 years in this situation, followed by death	**15 years** in this situation, followed by death	**10 years** in this situation, followed by death
ⓘ Physical health	**Very serious** problems	**Moderate** problems	**No** problems
ⓘ Living situation	**Very** satisfied	**Very** satisfied	**Very** satisfied
ⓘ Feeling useful	**Very** useful	**Very** useful	**Very** useful
ⓘ Making ends meet	**Reasonably** able to meet ends	**Reasonably** able to meet ends	**Very well able** to meet ends
ⓘ Receive support	**Dissatisfied**	**Dissatisfied**	**Very** satisfied
ⓘ Acceptance and resilience	**Able** to cope	**Not** able to cope **at all**	**Very well able** to cope
ⓘ Independence	Dependent	Independent	**Very** independent
ⓘ Mental health	**No** problems	**No** problems	**No** problems
ⓘ Social life	**Satisfied**	**Dissatisfied**	**Very** satisfied
	□	□	

- **Positive aspects are presented in light purple and negative aspects are dark purple**
- **Place your cursor on the ⓘ icon for descriptions of the aspects**

Which option do you prefer, B or C? (1/5)

		B	C
	15 years in this situation, followed by death	**15 years** in this situation, followed by death	**10 years** in this situation, followed by death
ⓘ Physical health	**Very serious** problems	**Moderate** problems	**No** problems
ⓘ Living situation	**Very** satisfied	**Very** satisfied	**Very** satisfied
ⓘ Feeling useful	**Very** useful	**Very** useful	**Very** useful
ⓘ Making ends meet	**Reasonably** able to meet ends	**Reasonably** able to meet ends	**Very well able** to meet ends
ⓘ Receive support	**Dissatisfied**	**Dissatisfied**	**Very** satisfied
ⓘ Acceptance and resilience	**Able** to cope	**Not** able to cope **at all**	**Very well able** to cope
ⓘ Independence	Dependent	Independent	**Very** independent
ⓘ Mental health	**No** problems	**No** problems	**No** problems
ⓘ Social life	**Satisfied**	**Dissatisfied**	**Very** satisfied
	□	□	

- **Positive aspects are presented in light purple and negative aspects are dark purple**
- **Place your cursor on the ⓘ icon for descriptions of the aspects**

Figure 9: Visual presentation of the pairwise choice task (translated from Dutch).

Conceptual framework and statistical analysis

In line with the conceptual framework of time-preference corrected QALY tariffs (see Jonker et al., 2018),[122] the utility derived by individual *i* for well-being state *j* in choice task *t* was defined as the product of the quality of life of the well-being state and the net present value (NPV) of the number of years lived in that well-being state, or:

$$U_{ijt}=\text{quality}_{ijt} * \text{NPV}(\text{years})_{ijt}+\varepsilon_{ijt} \tag{5}$$

An exponential discount function was used, which has a single discount rate parameter (*r*) that controls the degree of discounting and results in the following specification of the NPV:

$$\text{NPV}(\text{years})_{ijt}=(1-\exp\left(-r * \text{years}_{ijt}\right))/(\exp(r)-1) \tag{6}$$

The quality-of-life component in equation (1) was defined as follows:

Step 1.

$$\text{quality}_{ijt}=\beta_{i1}+\Sigma_{d=1}^{9}\,\beta_{i(d+1)} * \text{WOOPdomain}_{ijtd} \tag{7a}$$

in which β_i denotes a respondent-specific parameter vector that captures the importance of the nine WOOP domains (i.e. $\beta_{i(2\text{-}10)}$) relative to each other and to perfect well-being (i.e., excellent levels in all domains, captured by the β_{i1} intercept), and

Step 2.

$$\text{WOOPdomain}_{ijtd}=\Sigma_{L=1}^{5}\,\gamma_{dL} * X_{ijtdL}\ . \tag{7b}$$

in which γ_d denotes a WOOP domain-specific parameter vector that measures the relative importance of levels 2, 3 and 4 relative to levels 1 and 5 of each WOOP domain, subject to the constraints that $\gamma_{d1} \equiv 0$ and $\gamma_{d5} \equiv 1$ for identification, and where X_{ijtd} denotes a dummy-coded vector that equals 1 for the level at which each WOOP domain was presented to the respondent in the specific choice task, and 0 otherwise.

This specification was programmed in the BUGS language and fitted with OpenBUGS using Markov Chain Monte Carlo (MCMC) techniques. A technical appendix provides details about the statistical modelling. Worthy to note here is that the used approach implies by construction that the QALY decrements for levels 2 to 4 are monotonically increasing proportional to the γ_d WOOP domain-specific level importance parameters.

Results

A total of 2,660 respondents provided informed consent to participate in the study, of which 2,169 (82%) started with the DCE valuation tasks after the warm-up tasks. 2,012 respondents completed the full survey, which constitutes 93% of those who started with the DCE valuation tasks. The average age was 73 years, with 57% of respondents being male. The gender distribution of respondents above 75 years did not reflect the targeted sample quota, with females in this age category being underrepresented and males overrepresented (Table 8). Respondents generally reported high levels of well-being in the nine well-being domains of the WOOP (Figure 10). Lower levels were most frequently reported for the domains 'physical health', 'social life', 'feeling useful', and 'making ends meet'.

Table 8. Study sample characteristics (N=2,012).

	Sample	Sampling quota (census data)[1]
Male	57.3%	
Age in years (SD)[2]	73.3 (5.6)	
Age and gender distribution		
65-74 male	27%	26%
75+ male	30%	18%
65-74 female	28%	27%
75+ female	14%	29%
Finished tertiary education	35.4%	
Married	64.7%	
Employment		
Retired	84.9%	
Gainfully employed	6.0%	
Informal work and volunteering	5.6%	
Other	3.5%	
Country of birth		
Netherlands	94.2%	
Other	5.8%	
Cantril's Ladder (SD)	7.6 (1.2)	
EQ-VAS (SD)	73.4 (18.6)	

Note. SD, Standard deviation;[1] Data from Statistics Netherlands (Centraal Bureau voor de Statistiek) 2020; [2]Age ranged from 65 to 101.

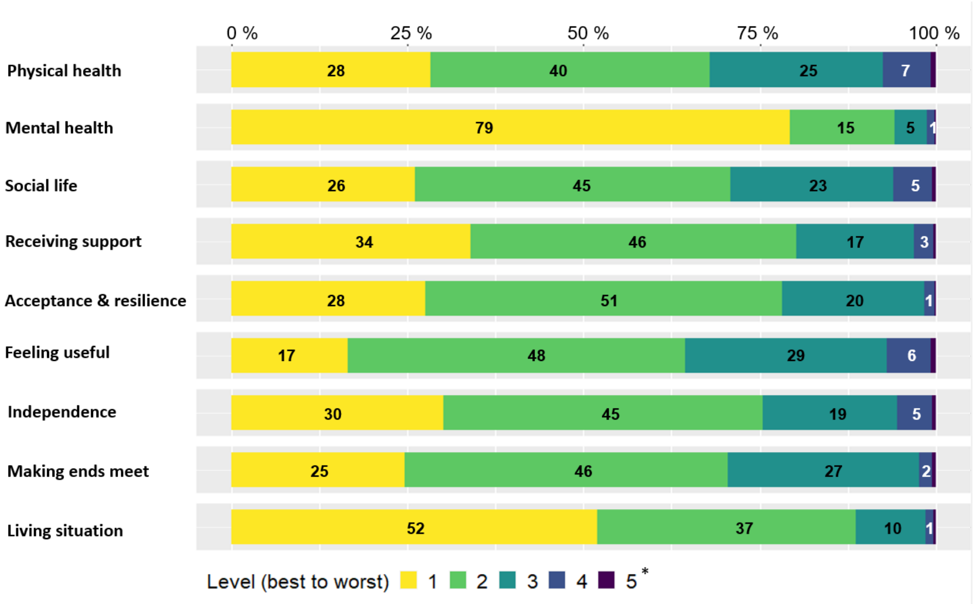


Figure 10: Distribution of responses to the nine well-being domains of the WOOP (N=2,012). *The worst level was selected by less than 1% of respondents in all WOOP dimensions

The average survey completion time was 34 minutes (median 24 minutes). Speeding, defined as a completion time of less than one-third of the median, occurred in 2% of responses (speeders were not excluded from the analysis). The cognitive debriefing questions in general provided favourable results, for instance, 78% of individuals at least partially agreed to the statement that the choice tasks were 'clear' to them (details in suppl. material, Table A1). We did not find large or significant differences in the response patterns to the cognitive debriefing questions between the three different experimental designs used (suppl. material, Table A2). This alleviates concerns about sacrificing (too much) respondent efficiency at the gain of statistical efficiency, which has been discussed before.[125,126]

Utility estimates

The calculated domain importance coefficients (equation 7a) show that 'physical health' and 'mental health', and to a lesser degree 'making ends meet', were the most important well-being domains among the older people in our sample (Table 9). Similarly, when summarising and plotting the terms used by respondents for describing well-being in their own words, physical and mental health were most frequently mentioned (see suppl. material, Figure A1).

Table 9. Domain importance on latent utility scale.

Domain	Mean	Lower 95% CI	Upper 95% CI	SD
Physical health	-1.381	-1.482	-1.283	0.051
Mental health	-1.507	-1.615	-1.401	0.055
Social life	-0.556	-0.606	-0.507	0.025
Receiving support	-0.457	-0.506	-0.409	0.025
Acceptance and resilience	-0.543	-0.596	-0.493	0.026
Feeling useful	-0.426	-0.475	-0.380	0.025
Independence	-0.718	-0.781	-0.657	0.032
Making ends meet	-1.136	-1.218	-1.054	0.042
Living situation	-0.674	-0.735	-0.615	0.031

CI = Credible Interval.

The anchored domain level utility weights are presented in Figure 11 (suppl. material, Table A3 shows the 95% CI). By construction, the estimated domain level weights are logically consistent within all nine well-being domains and non-positive. Two levels failed to reach statistical significance (i.e., the second-best levels of 'acceptance & resilience' and 'making ends meet'). The strongest decrements were found for 'mental health' (-0.329), 'physical health' (-0.302) and 'making ends meet' (-0.248), followed by 'independence' (-0.157) and 'living environment' (-0.147). 'Social life', 'receiving support', 'acceptance and resilience', and 'feeling useful' were generally perceived as less important for well-being. The theoretical spread of the WOOP utility ranges from -0.616 (worst possible state) to 1 (best possible state). The estimated discount rate was 0.173, considerably larger than has been found in a related general population study (0.057).[122] A higher discount rate may relate to a lower remaining life expectancy among older people. A previous study also found that people with more severe health problems had a higher discount rate,[122] thus finding this higher discount rate for older people, who tend to have more, and more severe health problems, is not completely unexpected.

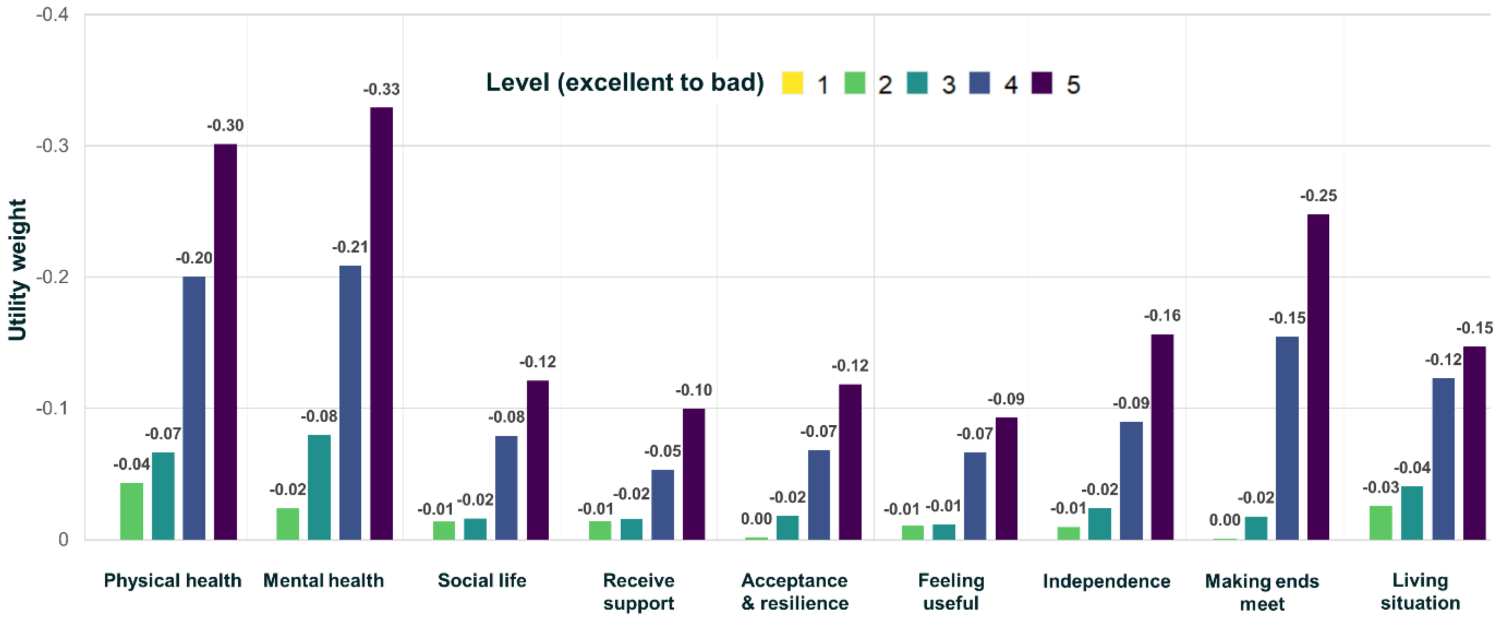


Figure 11: Utilities weights of the WOOP domain levels, with level 1 (excellent) as reference category.

Applying the utility tariffs to the WOOP responses in the sample produced a mean WOOP utility of 0.856 (SD 0.120). A utility value of 1 was observed for 34 respondents (1.7%) and a utility value below 0 (-0.067) for one respondent. The 25%, 50%, and 75% quantiles were 0.831, 0.889, and 0.929, respectively. When plotting utilities against EQ-VAS and Cantril's ladder (Figure 12), a strongly positive correlation was observed (r = 0.59 and r = 0.54, respectively) with similar trends for males and females.

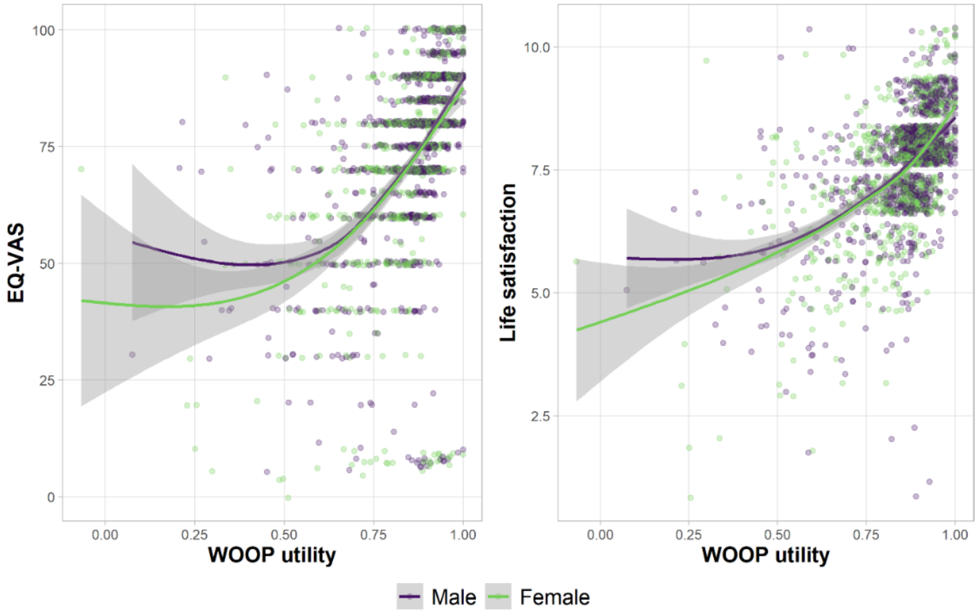


Figure 12: WOOP utility values plotted against health (EQ-VAS) and life satisfaction. For illustration purposes, jitter was added to the EQ-VAS and life satisfaction values, which are bounded on 0 to 100 and 0 to 10 range, respectively.

COVID-19 impact

Results for the question about whether the importance of the different well-being domains had changed due to COVID-19 were the following: The domains 'physical health', 'mental health', 'independence', and 'social life' generally appear to have become more important (suppl. material, Figure A2). Depending on the dimension, between 61% and 74% of respondents indicated that each respective domain had remained equally important for their well-being, with the lowest value observed for 'social life'.

Discussion

Given the increasing relevance of health and social care services for older people, and the fact that these services usually aim to improve well-being rather than health (alone), adequate measures for measuring the well-being of older people are required. The WOOP was recently developed for this purpose. To be useful as outcome measure in economic evaluations, such a measure ideally is accompanied by utility tariffs. Hence, in this study we present the results of a discrete choice experiment fielded among 2,012 individuals in the Netherlands aged 65 years and above to obtain preference-based utility tariffs for the WOOP. The resulting tariffs enable transformation of well-being states described with the WOOP into a utility score anchored on perfect well-being (1) and dead (0), and hence the use of the WOOP as outcome measure in cost-utility analyses of interventions in health and social care aimed at older people.

We elicited preferences from individuals aged 65 years and above, hence in the group of older people themselves and not in the general adult population as is commonly done for other outcome measures. Therefore, the utility tariffs for the WOOP reflect the relative importance for well-being of the different domains and functioning levels therein in the target population of the WOOP. This approach was deemed most relevant in informing the allocation of resources intended to improve the well-being of older people, and especially to evaluate optimal allocation *within* the budget for health and social care services for older people according to their preferences. Therefore, in contrast to measures like the EQ-5D, the WOOP is specifically targeted at one age group and not intended for comparisons across all adult age groups.

Given the large descriptive system and the target population of the WOOP, we conducted a pilot study to select the optimal elicitation method[127] and undertook several steps to reduce the cognitive burden of the choice tasks. Based on the responses to the cognitive debriefing statements presented to respondents after the choice tasks, it seems that the combination of a stepwise introduction to the experiment, colour-coding, level overlap and the separation of the trade-offs between well-being domains and duration was successful in reducing the cognitive burden to a manageable amount for this sample of older people.

In line with the Q-methodology study conducted to identify the domains of the WOOP[92] and with previous research,[128,129] we found that 'physical health' and 'mental health' were the most important domains for the well-being of older people, followed by 'making ends meet'. Domains like 'independence', 'social life', 'receiving support' and 'feeling useful' seem to be of somewhat lesser importance to their well-being. The relatively low importance of the domain 'social life' was somewhat surprising given the results of previous research.[92,128,129]

Strengths and limitations

Previous studies estimating utility tariffs for well-being measures primarily applied best-worst scaling (BWS) approaches[33,47,102] (or intend to do so[106]). Therefore, a noteworthy strength of the applied methodology is that it provides a feasible alternative approach, which was also shown to be preferable for older people in terms of the cognitive burden of choice tasks in a pilot study.[127] Moreover, the DCE design with a duration attribute allowed anchoring the utility weights of the WOOP on a QALY-like scale, facilitating a more straightforward combination of length and quality of life in computing the benefits of interventions. The applied approach furthermore accounts for non-linear time preferences, which otherwise would bias estimates in DCE with duration approaches.[122] The estimated discount rate of 0.173 implies that parameter estimates would have been severely biased by time preferences if we would have assumed linear time preferences. A more general implication of this is that for older people, estimated/empirical discount rates are much higher than the discount rates used in traditional HTA calculations, which mostly range between 1.5% and 5%.[130] The use of exponential (as opposed to, for example, hyperbolic) discounting in our analysis is also consistent with the common approach to discounting of health effects in health technology assessment.

More particularly, the implemented modelling approach has the advantage that it reduces the number of respondent-specific parameters, allows for correlated preferences between the WOOP domains, and produces readily available estimates of the relative importance of the WOOP domains, while ensuring a logically consistent utility tariff. The modelling approach used here was more structured than the one used by Jonker et al.,[122] but a more parsimonious model structure was crucial considering the large descriptive system of the WOOP and the limited number of respondents relative to the number of utility decrements.

While the pilot tests indicated that we reduced the complexity of the DCE choice tasks to a manageable cognitive burden for most respondents, decision heuristics could still have played an important role. For instance, while the colour-coding helped in identifying the differences between the two well-being states in a choice task, it may have stimulated respondents to focus on the colour intensity when making their choices. To what extent respondents used decision heuristics in general is unknown, although 89% of respondents reported to have compared all different aspects before making their choices (suppl. material, Table A1). At the same time, colour-coding combined with level overlap have previously been established as effective strategies to reduce the use of (other) decision heuristics.[99]

Reducing the complexity also entailed simplifying the attribute and level descriptions in the choice tasks—as is common in health state valuation.[32] We do not know whether and how this might have impacted the interpretation of attributes and levels, as we did not formally test the equivalence of abbreviated and full descriptions. Nevertheless, this is not expected to have had a substantial impact on the valuation results. First, as much as possible, all domains and levels were abbreviated in the same manner. Second, prior to the valuation tasks, individuals already were introduced to the full WOOP instrument with the full descriptions. Third, respondents had the full attribute descriptions available as mouse-over elements in the choice task to ease interpretation.

A clear limitation of the analysis relates to the representativeness of the sample, which is hampered by two factors. First, females aged 75 years and above were underrepresented in the sample (see Table 1). The market research company was unable to reach the desired number of completes in this group even after considerable effort.

Second, and more importantly, people above 65 years of age, who are part of online survey panels, and perhaps especially those above 75 years, likely will not be fully representative of this age group in terms of functioning, living situation, digital skills, and cognitive ability. Unfortunately, we did not collect data about these characteristics, but we could use EQ-VAS values as an indicator. When we compare age- stratified EQ-VAS values in our sample with data from a previous large-scale study among community dwelling Dutch elderly (suppl. Materials, Table A3), we find that individuals in our sample have a lower level of health.[131] This might hint towards capturing a wider range of respondents than just community dwelling individuals, but this cannot be confirmed. At the same time, it is very likely that people in poor health and well-being states are underrepresented in our study. For instance, another study found that among residents of nursing homes in the Netherlands, the mean EQ-5D VAS score was 64.8 (SD 21.7),[132] which is clearly lower than in our sample. The underrepresentation of individuals in poorer states, including those in nursing homes may explain the high levels of well-being in most domains of the WOOP observed in our sample.

Worth mentioning in this context is that 18% of survey participants dropped out during the introduction and warm-up tasks. It is likely that this drop-out is related to the cognitive capabilities of participants, which may have further contributed to an analysis sample of relatively capable, healthy, and happy respondents.

The preferences of older people in poorer states, including those that who are frail, dependent, or living in nursing homes, may thus differ from what we observed in our sample. These groups are, however, difficult to reach and experience more difficulty with participating in (this type of) research. As such, the utility tariffs presented here may not fully reflect the preferences of the older population in its entirety. Given the aim of the WOOP, assessing whether preferences regarding the WOOP states differ in the subgroup of the oldest old and frailest individuals is important, but appears to require a different study design. This might include purposive sampling, also within nursing homes, and interviewer-assisted survey techniques, with an adjusted and simplified choice experiment.

Finally, we emphasize that the data for this study was collected during the COVID-19 pandemic, an extraordinary context with special relevance to older peoples' well-being. Our attempts to assess the impact of this on the estimated preferences showed that 'physical health', 'mental health', and 'independence' domains may have especially increased in importance (see suppl. material, Figure A1). It is not clear whether possible effects of the pandemic on preferences for the WOOP domains are temporary or may last after the pandemic is over. After all, a possible effect of the current crisis may be that people became more aware of what they consider most important for their well-being and, hence, the preferences we measured in this study may even be closer to their true preferences.

Application and future research

The estimated utility tariffs enable the use of the WOOP in economic evaluations of health and social care interventions targeted at older people, first of all in the Netherlands. Using the WOOP may provide a more comprehensive overview of the benefits of such interventions as compared to health-related quality of life measures (e.g. EQ-5D), but also as compared to the ICECAP-O and the ASCOT.[81] Moreover, the WOOP has the advantage over other well-being measures that its utility tariff is anchored on dead and perfect wellbeing, facilitating a more straightforward combination of length and quality of life in computing the benefits of interventions. Consequently, the WOOP may also be useful when evaluation cross-sectoral interventions, for instance health and social care services combined with housing or income support. However, until further research confirms the (psychometric) validity of the WOOP and assessing interventions in health and social care in terms of their full benefits to older people becomes more established, we would advocate the use of the WOOP next to standard measures of health-related quality of life. This is also in line with the current recommendation of the Dutch health care institute (Zorginstituut Nederlands) for the use of the ICECAP-O. We do note that since the WOOP captures broader wellbeing including health, the measure cannot be readily added to results obtained using generic health-related quality of life measures, as this would imply double-counting.

Decision makers and analysts need to be aware that using an outcome measure like the WOOP, which focuses on broader outcomes than health and is conceptually targeted at a specific age group, makes it difficult to compare the results of evaluation studies with those using other outcome measures. Hence, the comprehensiveness and relevance of the WOOP in the specific context of health and social care for older people comes at the price of reducing the comparability of findings with those from economic evaluations in other populations or focused on health as outcome. Furthermore, for economic evaluations using the WOOP to be truly informative for decision-making about whether or not to implement particular health and social care services, a threshold value representing the monetary value of a well-being adjusted life year

(WALY) is required. Considering that the scope of benefits is broader, it is likely to be higher than that the threshold for a QALY. While different methods may be used to estimate such a threshold value,[38,133,134] an important conceptual question will be whether this valuation should be done within the target population, as the beneficiaries of health and social care interventions, or within the general public, as the payer of such interventions in a collective system (like in the Netherlands). A last noteworthy aspect of the use of broader outcome measures in general is that by extending the scope of the benefit dimension, one needs to consider also extending the cost dimension beyond health care to stay within a consistent framework.

Conclusion

By generating utility weights, the WOOP can now be used in economic evaluations of health and social care services targeted at older people, first of all in the Netherlands. Furthermore, the methodological approach used in this study may be helpful for future studies valuing newly developed measures with similarly large descriptive systems.

Appendix

Well-being of Older People measure (WOOP)
For each section, select the description that is most appropriate for you today.

Physical health

Consider physical conditions or ailments and other physical impairments that affect your daily functioning.

□ I have no problems with my physical health
□ I have slight problems with my physical health
□ I have moderate problems with my physical health
□ I have severe problems with my physical health
□ I have very severe problems with my physical health

Mental health

Consider problems with your ability to think, anxiety, depression and other mental impairments that affect your daily functioning.

□ I have no problems with my mental health
□ I have slight problems with my mental health
□ I have moderate problems with my mental health
□ I have severe problems with my mental health
□ I have very severe problems with my mental health

Social life

Consider your relationship with your partner, family or other people who are important to you. This concerns the amount and quality of the contact you have.

□ I'm very satisfied with my social life
□ I'm satisfied with my social life
□ I'm reasonably satisfied with my social life
□ I'm dissatisfied with my social life
□ I'm very dissatisfied with my social life

Receive support

Everyone needs help or support sometimes. Consider practical or emotional support, for example from your partner, family, friends, neighbours, volunteers or professionals. This concerns being able to count on support when you need it, as well as the quality of the support.

□ I'm very satisfied with the support I get, when needed
□ I'm satisfied with the support I get, when needed
□ I'm reasonably satisfied with the support I get, when needed
□ I'm dissatisfied with the support I get, when needed
□ I'm very dissatisfied with the support I get, when needed

Acceptance and resilience

Consider your acceptance of your current circumstances and your ability to adapt to changes to these, whether or not with support of your religion or belief.

□ I'm more than able to deal with my circumstances and changes to these
□ I'm able to deal with my circumstances and changes to these
□ I'm reasonably able to deal with my circumstances and changes to these
□ I'm not able to deal with my circumstances and changes to these
□ I'm not at all able to deal with my circumstances and changes to these

Feeling useful

Consider meaning something to others, your environment or a good cause.

□ I feel very useful
□ I feel useful
□ I feel reasonably useful
□ I do not feel useful
□ I do not feel at all useful

Independence

Consider being able to make your own choices or doing the activities that you find important.

□ I feel very independent
□ I feel independent
□ I feel reasonably independent
□ I feel dependent
□ I feel very dependent

Making ends meet

Consider having enough money to meet your daily needs and having no money worries.

□ I'm more than able to make ends meet
□ I'm able to make ends meet
□ I'm reasonably able to make ends meet
□ I'm not able to make ends meet
□ I'm not at all able to make ends meet

Living situation

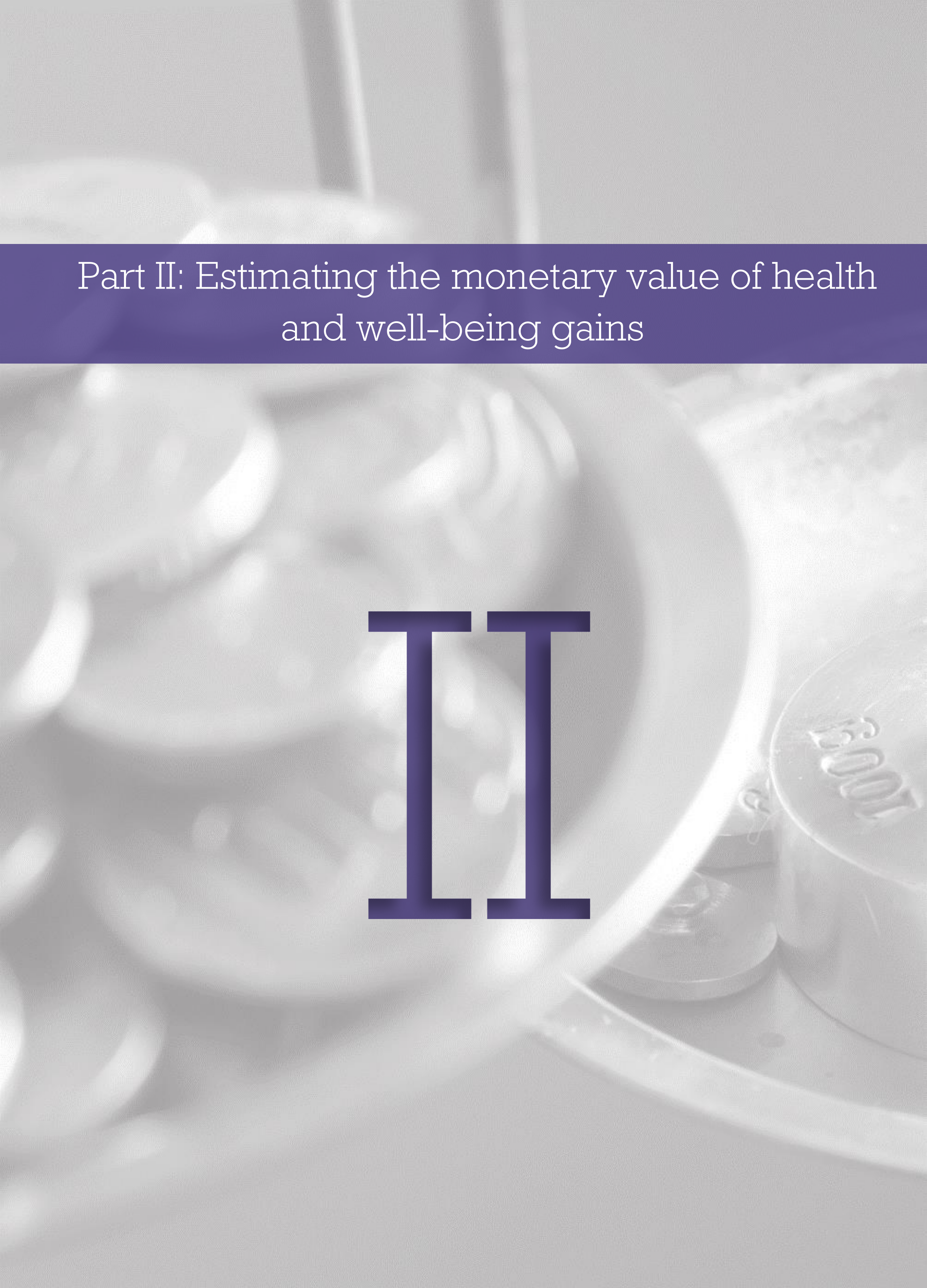
Consider living in a house or neighbourhood you like.

□ I'm very satisfied with my living arrangements
□ I'm satisfied with my living arrangements
□ I'm reasonably satisfied with my living arrangements
□ I'm dissatisfied with my living arrangements
□ I'm very dissatisfied with my living arrangements

Technical appendix

The specification described by equations 5 to 7b was programmed in the BUGS language and fitted with OpenBUGS using Markov Chain Monte Carlo (MCMC) techniques. This involved the selection of prior densities for the model parameters and updating these densities with the likelihood of the observed data. A multivariate normal prior was placed on the β_i parameters, i.e., $\beta_{\mathrm{i}} \sim \mathrm{MVN}(\mu, T)$. Uninformative normal priors (i.e., with means of 0 and standard deviations of 10) were assigned to μ and a Wishart prior with an identity scale matrix and 10 degrees of freedom to the precision matrix T. A uniform (0,1) prior was placed on r, and Dirichlet priors with concentration parameters equal to 1.0 were assigned to a set of latent $\gamma^*_{d(1:4)}$ parameters that were subsequently transformed into $\gamma_{d(1:5)}$ by setting $\gamma_{\mathrm{d1}} \equiv 0$ and defining $\gamma_{dl} = \sum_{m=1}^{l} \gamma^*_{dm}$ for $l \in$ 2-5. This ensured, by construction, monotonically increasing $\gamma_{d(1:5)}$ parameters that automatically adhered to the required $\gamma_{d1} \equiv 0$ and $\gamma_{d5} \equiv 1$ constraints, leading to monotonically increasing level-importance parameter estimates within each of the WOOP domains.

Standard Gibbs updates were used to update μ and Σ, antithetic Metropolis-within-Gibbs update steps to update β, slice sampling update steps to update r, and non-conjugate random-walk Dirichlet update steps were used to update the γ^* parameters. In addition, the implied decrements for the WOOP attribute levels on the QALY scale were calculated by first dividing all elements in the mean vector (μ) by the first element (μ_1), which ensured that the value of full well-being was equal to 1.0, and then, for each WOOP domain, multiplying the scaled average domain importance (i.e. $\mu_{2\text{-}10}/\mu_1$) with the corresponding domain-specific γ_d parameter. Given the constraints on $\boldsymbol{\gamma_d}$, this implies, by definition, that QALY decrements for level 1 are 0 and that the QALY decrements for levels 5 are equal to $\mu_{(d+1)}/\mu_1$ and thereby equal to the relative importance of the respective WOOP domain. Furthermore, this implies that the QALY decrements for levels 2 to 4 are monotonically increasing proportional to the γ_d WOOP domain-specific level importance parameters.



Part II: Estimating the monetary value of health and well-being gains

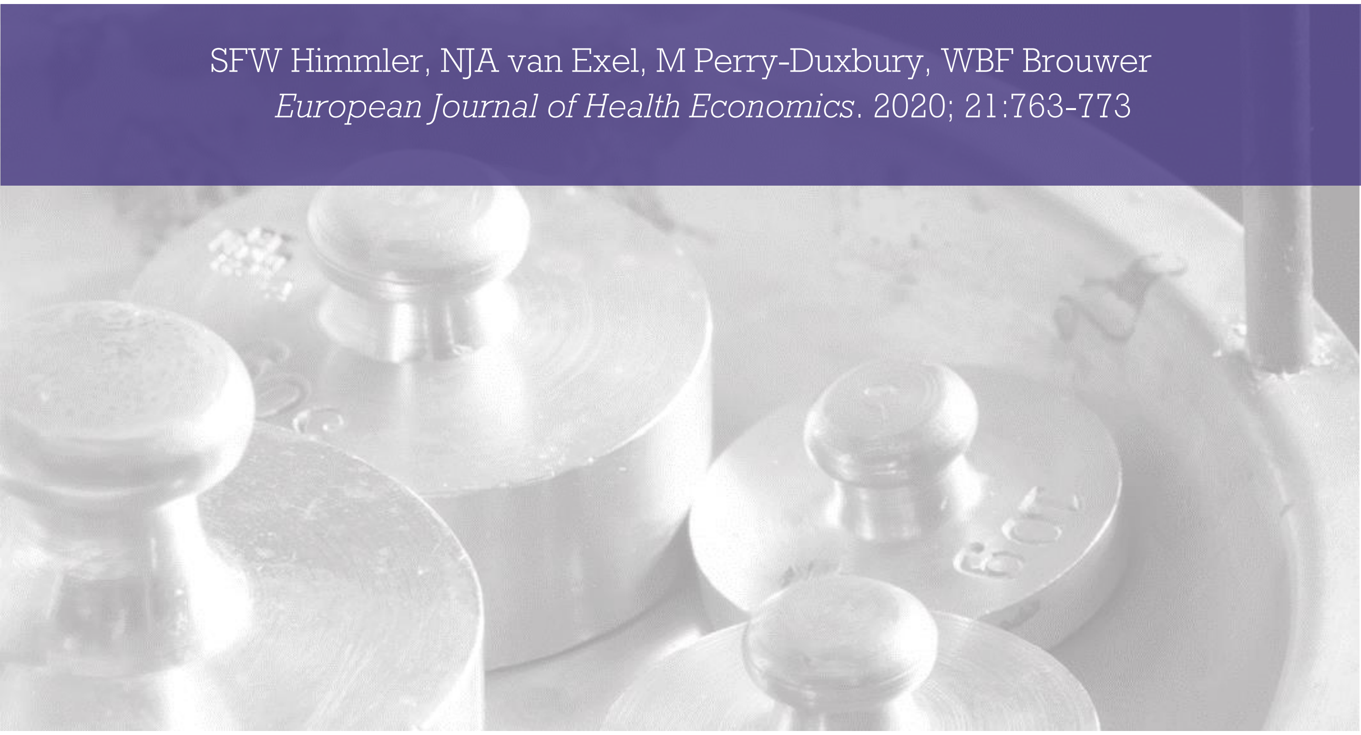
II

5

Willingness to pay for an early warning system for infectious diseases

SFW Himmler, NJA van Exel, M Perry-Duxbury, WBF Brouwer

European Journal of Health Economics. 2020; 21:763-773



Abstract[*]

Early warning systems for infectious diseases and foodborne outbreaks are designed with the aim of increasing the health safety of citizens. As a first step to determine whether investing in such a system offers value for money, this study used contingent valuation to estimate people's willingness to pay for such an early warning system in six European countries. The contingent valuation experiment was conducted through online questionnaires administered in February to March 2018 to cross-sectional, representative samples in the UK, Denmark, Germany, Hungary, Italy, and The Netherlands, yielding a total sample size of 3,140. Mean willingness to pay for an early warning system was €21.80 (median €10.00) per household per month. Pooled regression results indicate that willingness to pay increased with household income and risk aversion, while they decreased with age. Overall, our results indicate that approximately 80–90% of people would be willing to pay for an increase in health safety in the form of an early warning system for infectious diseases and food-borne outbreaks. However, our results have to be interpreted in light of the usual drawbacks of willingness to pay experiments.

[*] The online version of this article (https://doi.org/10.1007/s10198-020-01171-2) contains supplementary material relating to appendices A to F in this chapter.

Introduction

Increasing the health safety of citizens is an important policy goal in countries across the world. Recent infectious outbreaks of, for example, Ebola, SARS, bird flu, and salmonella, emphasise that improving safety cannot always be realised by countries separately.[135] Recently, for example, the European Union has initiated an interdisciplinary research network that investigates the potential for an international, integrated early warning system for identifying, containing and mitigating large infectious outbreaks more rapidly (*http://www.compare-europe.eu/*).

Establishing and maintaining such a system would likely entail considerable costs. To determine whether this would be money well spent, it is essential to consider all its potential benefits. The relevant benefits could include a reduction in disease burden, increased feeling of safety, or the mitigation of economic consequences of infectious diseases and food-borne outbreaks, which can be considerable for countries, organisations, and individuals. For instance, the economic impact of the Ebola crisis in 2014-2015 on Sierra Leone, Guinea and Liberia was estimated at $2.8 billion.[136]

However, in general, reliable evidence and estimates of these potential benefits of an early warning system, separately or overall, are scarce and difficult to obtain, especially in the case of multinational initiatives. In light of this and the fact that the full potential benefits would include, besides aspects like health gains, also elements like improved feeling of health safety, it is not possible to quantify the overall benefits of such an international early warning system based on existing data.

Therefore, in this study, we aim to provide an indication of the *perceived* overall value of such a system in terms of improving citizen's feelings of health safety. For that purpose, we first develop a contingent valuation willingness-to-pay approach, which provides such a valuation, *given* beliefs and sentiments in the population regarding all different aspects of a warning system. Second, we apply this approach in six selected countries across Europe (i.e., Denmark, Germany, Hungary, Italy, the Netherlands, and the UK) to derive a range of estimates and assess the potential implications of our results on an international level.

This paper summarises our efforts to accomplish these goals and its remainder is divided into four sections. First, we briefly summarise the findings from a previous literature review surrounding the methods that have been applied in similar contexts, namely valuing health safety, to motivate the chosen approach further. After that, we consecutively report on the design and administration of our experiment, the data analysis, present the results of our study, and conclude the paper with a discussion of the limitations and implications of our findings.

Background

The introduction of an international integrated warning system to increase health safety would not be necessary if communicable or infectious diseases were not a significant factor in the Global Burden of Disease. The Burden of Communicable Diseases in Europe project found an average disease burden in Germany alone of

33,116 Disability Adjusted Life Years (DALYs) per year for influenza and 19,115 DALYs per year for salmonella.[137] On a European level, influenza was estimated to be responsible for 81.8 DALYs lost per 100,000 population between 2009 and 2013, corresponding to 412,673 DALYs using the EU population size from 2011.[138]

Considering these substantial effects of infectious diseases, some of the potential benefits of an international integrated warning system become clearer. Of course, the real benefits also depend on the translation from warnings to effective interventions that prevent or mitigate the consequences of outbreaks. Besides possible health gains resulting from this, there are also less tangible benefits from having a warning system, which include an increase in health safety and feeling more secure. The valuation of these benefits may be less straightforward than calculating potential DALYs averted.

The valuation of interventions affecting safety is relevant both within and outside the health care setting. For example, environmental and transportation research is concerned with interventions, which aim to improve the safety of recipients. Perry-Duxbury et al. conducted a literature review in which they examined the methodologies of empirical research valuing safety from all relevant fields, including environment, transportation and health.[139] Of the 33 papers reviewed, 22 were found to use the contingent valuation method to value the effects of safety-affecting interventions. The four papers in the field of health that empirically valued interventions increasing health safety, all used a form of stated preference methodology. These papers aimed to estimate the value of reducing mortality risks,[140] preventing child maltreatment deaths,[141] reducing the risk of sexually transmitted diseases,[142] and vaccinations in pandemic outbreaks.[143] The first three papers used willingness to pay (WTP) contingent valuation method, while the last paper used a discrete choice experiment to elicit valuations.

The literature review identified income to be a significant predictor of WTP in all included contingent valuation studies.[144] A higher level of education was associated with a higher WTP in six of the nine papers that included information on education. Age and gender both also had strong correlations with WTP. However, these correlations were positive in some of the studies and negative in others. The literature review also reported results regarding relationships of WTP with risk (perception). For example, individuals that had been directly or indirectly exposed to the outcome of interest reported a higher WTP, as did those who had a higher level of perceived risk, were more knowledgeable or more concerned about the issue, or were more concerned than others about the outcome under study. Finally, study design elements were shown to affect WTP estimates. For example, presenting scenarios with higher baseline risk was associated with a higher WTP. In addition, different studies found that presenting higher intervention costs or more information about the intervention in the scenario description also affected the estimated WTP. However, the direction of the effect differed between studies. The information provided by the literature review guided some of the methodological choices of our study, which are described next.

Methods

Survey administration and piloting

To estimate the WTP for an international integrated early warning system for infectious diseases and food-borne outbreaks, we conducted contingent valuation experiments utilising general population samples from six European countries: Denmark, Germany, Hungary, Italy, the Netherlands, and the UK. Sampling and administration of the WTP questionnaire were conducted by a professional sampling agency, from February to March 2018, using an online survey format. The sampling agency recruited participants from existing online panels. The survey was administered to citizens aged between 18 and 65. Individuals aged 65 and above were not included for two reasons: First, recruiting elderly respondents from online panels can be challenging in some of the included countries. Second, we wanted to limit our population to the (income) taxpayers, as we used a tax increase as payment vehicle in the experiment. The samples were aimed to be representative for national populations regarding age, gender, and level of education, with a sample size of around 500 individuals per country. Participants were able to complete the questionnaire on a computer or mobile device. They did not receive a personal financial reward for engaging in the experiment but could choose a charity, which would receive a small donation after completing the survey. Participants had to consent to their information being used for research purposes and were free to drop out of the experiment at any time.

The reasoning behind the country selection was to cover a variety of cultural perspectives relevant to the valuation of safety and public intervention. The latter was assessed by applying the three most relevant dimensions of Hofstede's cultural dimensions theory in this context: individualism vs collectivism, masculinity, and uncertainty avoidance.[145] The included countries furthermore constitute a mix of different levels of social and economic development in Europe. The questionnaire, which was initially developed in English, was translated into Danish, German, Hungarian, Italian, and Dutch by professional translators and checked for comprehensibility and consistency by native speakers. In designing the experiment and payment scales, GBP and EUR values were assumed to be equivalent, while monetary values and payment scales were converted from GBP into DKK and HUF using the mean exchange rate from February 2018. In the case of Hungary, this was additionally adjusted for purchasing power.[146] Payment scales were rounded to natural integer values in all survey versions to prevent peculiar payment options. The payment scale of the UK survey and the equivalent monetary values for Danish crowns and Hungarian forint can be found in online Appendix D.

Before the launch of the main survey, the questionnaire was tested in both a group of experts in infectious diseases and food-borne outbreaks associated with the COMPARE research network (n=22) and a representative sample of the public in the UK (n=134) in January 2018. The length of the survey was slightly reduced following the pilot tests. After this stage of piloting, the questionnaire was fielded in a representative sample of UK citizens (n=533). To test the payment scale used in the experiment, we administered two additional surveys (n=500 each): One with smaller payment options, and the other asking for yearly contributions instead of monthly contributions. The validity of the results of the three survey versions was assessed based on whether WTP was influenced by income and based on a comparison to a reference point (home contents insurance, a common type of insurance, which covers the possessions in your home against risks like fire, theft, and storm), which was included in the surveys. The initial payment scale performed best and was therefore used in all surveyed countries.

Survey design

The general design of the WTP experiment followed the structure of an existing survey, which was purposely designed to elicit the WTP for a quality-adjusted life year (QALY).[147] After a brief introduction to the topic at hand and the purpose and design of the questionnaire (see online Appendix A), respondents had to state their age and gender before describing their current health using a generic health instrument (EQ-5D-5L).

The following part of the questionnaire started with a "warm-up" WTP exercise, where participants had to state their WTP for a pair of shoes. This elicitation task was included to familiarise respondents with the procedure and to test whether the chosen approach resulted in reasonable estimates for a common market good. Next, respondents started with the central WTP task: valuing the early warning system. A two-stage procedure consisting of a two-step payment scale approach and an open-ended question was applied to elicit individuals' WTP. The motivation for this approach has been outlined elsewhere.[147–149] In summary, it intends to provide precise and direct maximum WTP valuations, using a stepwise procedure that helps respondents to form and articulate their preferences.

The scenario outlined to respondents was that establishing and maintaining an international integrated warning system, which could contain and mitigate infectious disease and food-borne outbreaks, naming Ebola, SARS, bird flu and salmonella as examples, is not without costs. Participants then were asked to imagine that the funding of such an international warning system would take place through national taxation in the participating countries. All eligible people in their country (aged 18 and above) would have to contribute via monthly instalments starting immediately. The payment was framed as a recurrent tax since most respondents in European countries are likely familiar with similar forms of payments. This scenario did not include information on the magnitude of the potential health benefits. The reasoning behind that was to emphasise the perceived feelings of health safety due to such a system in the elicitation rather than a particular hypothetical gain in health, also because the potential benefits are uncertain at this stage. This could provide a broad valuation based on the beliefs and attitudes of the respondents themselves. Information on the types of *local* systems already in place and how these would be integrated into this international system was also omitted. Not only would this be cognitively burdening, but the chosen approach also conformed more closely to the general definition of the

COMPARE project and hence warning system at this stage. While this leaves respondents with imperfect information, this was intentional, as our goal was to value a warning system which features are not yet fully clear, in terms of the incremental feelings of health safety that comes with it.

In the first step of the initial stage of the willingness-to-pay experiment, respondents were asked to indicate the amounts they would definitely be willing to pay per month for having this international, integrated warning system, using a payment scale ordered from low to high GBP or EUR values (0, 1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 100, 120, 150, 200, more). The payment scale for the UK contained the same values in GBP, the version for Denmark contained the same values converted to DKK (and rounded), and the version of Hungary was adjusted for purchasing power and converted to HUF (and rounded). The Hungarian and the Danish scale are included in online Appendix D.

Individuals who chose the "more" option on the payment scale subsequently had to indicate a value higher than 200 in an open-ended question. Individuals who chose 0 as their maximum WTP had to select one of the following options to specify the reason for this answer, with the following predefined options: (i) not worth more than 0, (ii) unable to pay more than 0, (iii) government task, or (iv) the option to formulate another reason in an open text field. The former two options were considered to indicate a true WTP of zero, while "government task" was designated as a protest zero. Entries in the open text field were evaluated and labelled as either true zero or protest zero. Individuals who chose a value between 1 and 200 were subsequently asked to mark the amounts they would definitely *not* be willing to pay per month on the same payment scale, excluding the WTP values they had selected in the preceding step.

Jointly, these two steps generated a WTP interval between the highest amount that a respondent definitely was willing to pay and the lowest amount he or she was definitely not willing to pay. In the second stage of the WTP procedure, respondents had to indicate an exact amount within this interval that was closest to the maximum that they would be willing to pay per month. Respondents could specify decimals in this second stage, not limiting the WTP to integer values. The elicited WTP amounts in the second step were taken as the best approximation of people's WTP for the (health safety benefits from an) international integrated early warning system for infectious diseases. Throughout the two steps, participants were reminded to keep their ability to pay in mind (their net monthly household income) before indicating any interval or specific value to prevent ex-ante mitigation.[150] The design and the exact wording of the WTP questions can be found in online Appendix C. The questionnaire continued with two additional WTP valuation scenarios involving different degrees of risk reduction and disease severity, which will not be discussed in this paper. Subsequently, respondents had to provide further socio-demographic information. Estimates for household income were obtained in a two-step process. Respondents first selected an income range before indicating an exact amount. Missing exact income amounts were imputed based on the sample means of the income interval selected in the first step, if applicable.

Respondents were furthermore asked about whether they or their family had ever been exposed to an emerging infectious disease or outbreak (yes/no), and about their general awareness related to emerging infectious diseases and food-borne outbreaks, which was queried using 12 statements and a 7-point Likert scale. The statements comprised of a collection of aspects found to be relevant in this context based on the findings from the literature review (see online Appendix B). Finally, respondents completed a brief version of the health-risk attitude scale (HRAS),[151] which consists of six statements about resolving risky health decisions that need to be ranked on a 7-point Likert scale ranging from "totally disagree" to "totally agree".

The survey ended with a module asking respondents whether they had home contents insurance, the size of the corresponding yearly premiums and how they would value the described early warning system in comparison to their contents insurance (lower, roughly the same or higher). These results of this final module were intended to serve two purposes: First, they were used to test different types of payment scales before the rollout of the main survey. Second, comparing the contents insurance premiums people actually pay and the stated relative value of early warning system and contents insurance serves as a validity check of the stated WTP values. In addition to the survey data, we collected country aggregate estimates on the relevant dimensions of Hofstede's cultural dimension theory (masculinity, individualism, and uncertainty avoidance) and the level of trust in public institutions.[152]

Data analysis

Before analysing the data, we converted all monetary values from Danish, UK, and Hungarian respondents to Euro values using the average exchange rates during the month of sampling (7.45 DKK/€, 1.14 £/€, 312 HUF/€). In the next step, cross-country data validity and comparability were assessed by exploratory, descriptive analysis. We first inspected the proportions of and reasons for zero WTP answers, distinguishing between true and protest zeros. We excluded protest zeros and WTP outliers from the remainder of the analysis. The latter was defined as WTP values larger than 5% of monthly household income. Descriptive statistics were calculated based on the remaining WTP valuations.

Linear regression analysis was conducted on the WTP valuations from all six countries to examine which factors influenced the WTP answers and whether the observed effects were in line with theoretical considerations as well as previous empirical findings of WTP determinants (outlined in the beginning of this chapter). The regression analysis thus functions as a validity check for our experimental design and WTP results. We also explored the suitability of Tobit or Two-part-models for the regression analysis, however using root mean squared error and mean absolute error as performance criteria revealed that standard linear regression provided the best model fit. Calculations were conducted using the pooled total sample, as well as the separate country-level samples. Descriptive analysis and regression analyses were performed using STATA 15.0 (Stata Corp. 2018. Stata Statistical Software: Release 15. College Station, TX: Stata Corp LP).

Results

Characteristics of country samples

The total number of completed surveys from the six chosen European countries was 3,140. Unfortunately, information on the response rate or the share of respondents starting, but not finishing the survey could not be obtained from the sampling agency. On average, it took respondents 18.9 minutes (SD 11.2) to complete the questionnaire. The six samples were well balanced regarding age, gender, and education in their respective countries for the aimed subset of individuals aged between 18 and 65. Descriptive statistics of the respondents per country are shown in Table 10. The average gross monthly household income ranged from €1,214 in Hungary to €6,417 in Denmark. Employment status and educational attainment varied between countries, as to be expected. The sub-samples also differed considerably in the rate of past exposure to infectious diseases and food-borne outbreaks (10% in the UK vs 62% in Hungary).

Zero responses and protest answers

Overall, 14.8% of respondents stated a WTP of zero, with a share of 7.3% in Italy at the lower end and 23.2% in Hungary at the upper end. Of those with a WTP of zero, most respondents chose the pre-specified option "Government task" (57.3%) and only to a lesser extent the options "Not worth it" (17.2%) and "Unable to pay" (15.3%) to justify a WTP of zero, with considerable differences between countries. Of the 47 qualitative responses in the category "Other", 40 were classified to be similar to "Government task" or as protest answers. The remaining seven qualitative responses were more related to whether the system would be worth installing. These, therefore, were included in the "Not worth it" category, which, together with "Unable to pay" category, represent true zeros. The entirety of "Government task" and further protest answers (N=306) were treated as protest zeros and therefore not included in the following WTP estimates and regression analysis. Table 11 presents the share of zero values per country as well as the indicated reasons for the zero valuations. The share of protest zeros among zeros varied between 53.5% in the UK and 78.6% in Hungary. Individuals who provided protest answers had a significantly lower income (p=0.010), higher age (p<0.001), lower level of education (p=0.046) and only little awareness of outbreaks (p<0.001) in comparison to respondents with non-protest answers.

Table 10: Descriptive statistics (SD in brackets).

	UK	DK	GER	HUN	IT	NL	Total
Monthly household inc. €[a]	3,339	6,417	3,076	1,214	2,495	2,715	3,214
	(2,974)	(9,004)	(1,919)	(1,149)	(1,662)	(1,632)	(4,372)
Age	42.06	40.99	43.08	41.76	41.65	43.52	42.18
	(13.65)	(14.55)	(13.35)	(13.23)	(13.94)	(14.91)	(13.97)
Female	0.50	0.49	0.52	0.51	0.52	0.49	0.51
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
No finished sec. education	0.02	0.08	0.02	0.03	0.02	0.03	0.03
	(0.15)	(0.28)	(0.15)	(0.16)	(0.12)	(0.16)	(0.18)
Finished high school	0.50	0.54	0.65	0.55	0.60	0.59	0.57
	(0.50)	(0.50)	(0.48)	(0.50)	(0.49)	(0.49)	(0.50)
Tertiary education	0.48	0.38	0.33	0.42	0.39	0.38	0.40
	(0.50)	(0.49)	(0.47)	(0.49)	(0.49)	(0.49)	(0.49)
Married [a]	0.60	0.52	0.58	0.62	0.57	0.57	0.58
	(0.49)	(0.50)	(0.49)	(0.49)	(0.50)	(0.49)	(0.49)
Employed	0.56	0.49	0.58	0.66	0.44	0.52	0.54
	(0.50)	(0.50)	(0.49)	(0.48)	(0.50)	(0.50)	(0.50)
Self-employed	0.09	0.06	0.10	0.08	0.19	0.08	0.10
	(0.29)	(0.24)	(0.30)	(0.27)	(0.39)	(0.27)	(0.30)
Unemployed	0.06	0.08	0.04	0.04	0.10	0.06	0.06
	(0.24)	(0.27)	(0.21)	(0.19)	(0.29)	(0.24)	(0.24)
Homemaker	0.11	0.03	0.07	0.04	0.09	0.06	0.07
	(0.31)	(0.16)	(0.26)	(0.19)	(0.29)	(0.24)	(0.25)
Student	0.06	0.17	0.08	0.07	0.10	0.11	0.10
	(0.23)	(0.37)	(0.27)	(0.25)	(0.30)	(0.31)	(0.30)
Retired	0.08	0.13	0.12	0.10	0.08	0.06	0.09
	(0.27)	(0.33)	(0.32)	(0.30)	(0.27)	(0.23)	(0.29)
Unable to work	0.05	0.05	0.01	0.02	0.00	0.11	0.04
	(0.21)	(0.22)	(0.12)	(0.15)	(0.06)	(0.31)	(0.20)
EQ-5D-5L sum score (0-100)	86.76	83.86	85.41	88.99	87.50	88.94	86.91
	(18.05)	(17.99)	(16.98)	(14.49)	(14.64)	(14.14)	(16.24)
Awareness of outbreaks [b]	52.89	50.91	51.64	52.68	55.15	49.95	52.21
	(8.06)	(7.79)	(8.51)	(8.21)	(8.04)	(8.26)	(8.31)
% no past exposure	0.90	0.67	0.72	0.38	0.87	0.69	0.71
	(0.30)	(0.47)	(0.45)	(0.49)	(0.33)	(0.46)	(0.45)
% no family past exposure	0.06	0.18	0.13	0.17	0.05	0.11	0.11
	(0.23)	(0.38)	(0.34)	(0.37)	(0.22)	(0.31)	(0.32)
% no personal past exposure	0.06	0.21	0.18	0.48	0.09	0.23	0.20
	(0.23)	(0.41)	(0.38)	(0.50)	(0.28)	(0.42)	(0.40)
HRAS [d]	29.32	27.17	28.87	28.68	30.10	28.83	28.84
	(5.99)	(5.55)	(5.92)	(4.89)	(5.32)	(5.88)	(5.68)
Observations	553	514	522	504	523	524	3,140

Note: [a] Includes registered partnerships or cohabiting; [b] from 12 to 84 (12 questions with 7 levels); [d] Health Risk Attitude scale from 6 to 42 (6 questions with 7 levels).

Table 11: Percentage of responses with WTP of zero

	% Share of zeros	"True zero WTP"		"Protest zero"
	(total)	Not worth it	Unable to pay	Gov't task + protest
UK	12.8	31.0	15.5	53.5
Denmark	11.9	23.0	19.7	57.3
Germany	15.7	20.7	15.9	63.4
Hungary	23.2	9.4	12.0	78.6
Italy	7.3	21.1	15.8	63.2
Netherlands	18.1	15.8	15.8	68.4
Total	14.8	18.8	15.3	66.0

Table 12: WTP per month in EUR excluding protest zeros and outliers[a]

	Mean	SD	Median	Min	Max	N
UK	20.74	32.63	9.11	0.00	284.80	496
Denmark	28.33	42.43	13.42	0.00	460.98	473
Germany	21.01	30.27	10.00	0.00	250.00	457
Hungary	8.89	13.80	3.85	0.00	144.21	397
Italy	27.32	33.05	15.00	0.00	202.00	457
Netherlands	22.71	29.04	10.00	0.00	250.00	433
Total	21.80	32.32	10.00	0.00	460.98	2,713

Note: [a] Outliers defined as WTP exceeding 5% of monthly income.

Outliers and willingness to pay estimates

Turning to the actual WTP estimates, the elicited values for the lower interval of the first stage of the WTP exercise ("definitely be willing to pay") had a mean of €14.68 (SD 23.65). The corresponding mean for the upper interval ("definitely not willing to pay") was €42.63 (SD 67.15). The second stage produced a mean stated WTP for an international integrated early warning system for infectious diseases and food-borne outbreaks of €25.17 (median €10.07) per month per household. The standard deviation of €42.87 exemplifies a considerable heterogeneity in WTP within and across countries.

Several outliers with values up to €1,000 per month influence the mean WTP. The proportion of respondents with a WTP above €100 in the analysed sample was 5.0% and ranged from 0.7% in Hungary to 8.8% in Italy. Some of these outliers might represent the real WTP of respondents, while others may be deliberate or incidental overstatements. Applying the above-described criterion, 4.8% of responses qualified as outliers (N=121) and were excluded from the remainder of the analysis. Doing so reduced the mean monthly WTP from €25.17 to €21.80 in the remaining sample of 2,713 observations. Table 12 presents the corresponding values and further summary statistics, while. Figure 13 presents the distributions of all WTP values on country level. For readability, values over €100 (4 % of the total sample) are trimmed off. The mean monthly WTP varied from €8.89 in Hungary and €28.33 in Denmark.

Results from the included reference point, home contents insurance, revealed that for 51.1% of insurance holders (68.9% had this type of insurance) the perceived value of the warning system was more or less equal to the value of the contents insurance. In the subgroup that provided information on their monthly premiums, the mean difference between WTP and stated insurance premium was €5.28 (50.7% within a €10 range). A higher perceived value of the warning system (24.7%) coincided with a WTP, which was larger than the insurance premium in 56.6% of cases. A lower perceived value (24.3%) fell in line with a relatively lower WTP in 56.5% of cases.

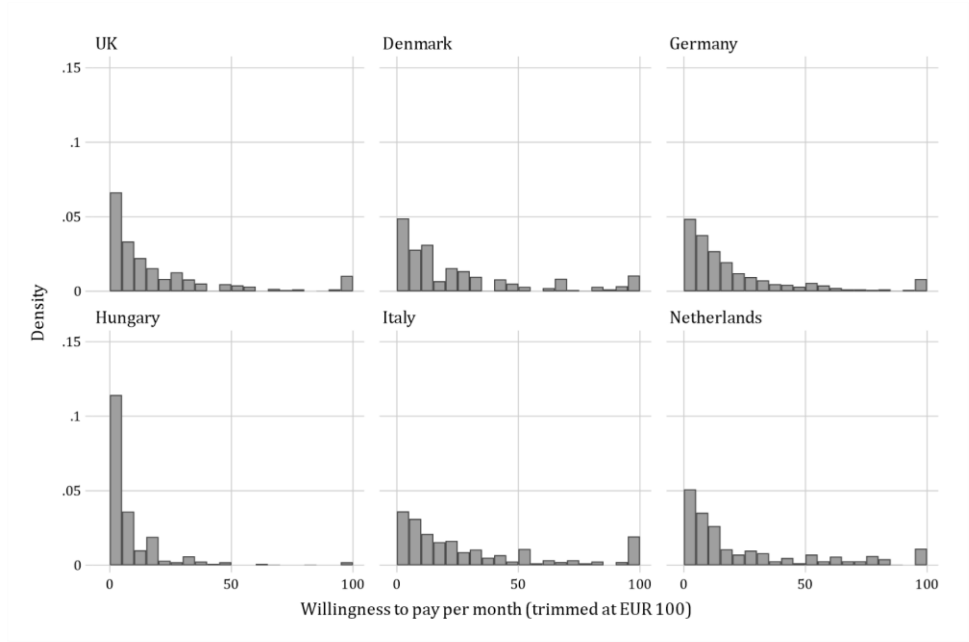


Figure 13: WTP values per country

Determinants of willingness to pay

Table 13 column one lists the results of regressing the WTP values on multiple individual characteristics using the pooled data from all six countries, excluding protest answers and outliers. To account for the correlation of errors within countries, we used cluster-robust standard errors on country level in the regression models. The number of observations dropped from 2,713 to 2,583, as some respondents did not provide any information on their household income. As the WTP data were skewed, we also analysed the data using log transformed WTP values. However, here we present the results using the raw WTP values as the general results, and implications of both approaches were highly similar. Moreover, the linear specification avoided having to drop zero WTP values and provides a more straightforward interpretation. The Log WTP results can be made available upon request.

Income had a highly significant and positive non-linear effect on the WTP, while age significantly reduced the WTP. Education did not affect WTP. The highest levels of awareness of outbreaks and health risk aversion (HRAS) seemed to influence WTP, although the coefficient of the former was not significant. Past exposure, marital status, or not being employed, did not significantly affect WTP.

The remaining columns of Table 13 present the results on country level. Factors affecting WTP differed considerably between countries with some coefficients even switching signs. Household income significantly increases WTP in all six countries, whereas age was significantly negatively associated with WTP in three of the countries. Consistently positive (but not always significant) coefficients were found for the highest quartiles of outbreak awareness and HRAS, i.e., being relatively most aware of the associated risks and being relatively most health risk-averse in general. Better health was associated with lower WTP throughout all countries. Alongside the differences in coefficients, the explanatory power of our model changed substantially between countries. The R-squared varied between 0.117 for the German model and 0.247 for the Italian model. Differences in model fit as measured by AIC/BIC and RMSE were even more substantial.

When including variables in a stepwise procedure, the conclusions for the pooled regression were reasonably stable across model specifications (see online Appendix E). Adding country dummy variables to the pooled model slightly diminished the effect of income. The respective coefficients of Hungary, Italy and the Netherlands were significant compared to the UK as reference category. This result indicates that even after controlling for socioeconomic characteristics, including income, WTP significantly differed between countries. Hofstede's cultural dimensions masculinity, individualism, and uncertainty avoidance, as well as trust in public institution further explained these differences (see online Appendix E).

Table 13: OLS regression on WTP excluding WTP outliers and protest zeros

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Pooled	UK	DK	GER	HUN	IT	NL
log income	10.0***	7.95***	14.44***	8.01***	5.55***	10.1***	8.08***
	(0.54)	(2.34)	(4.58)	(2.37)	(1.77)	(2.38)	(2.65)
age	-0.94**	-1.29*	-1.04	0.23	-1.27**	-1.71**	-0.27
	(0.29)	(0.73)	(0.92)	(0.71)	(0.53)	(0.85)	(0.69)
age-squared	0.01	0.01	0.00	-0.01	0.01**	0.02	-0.00
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
female	-3.85	-3.87	-13.4***	-6.28**	-0.48	0.67	0.67
	(2.38)	(2.95)	(3.50)	(2.78)	(1.24)	(3.25)	(3.16)
tertiary educ	2.41	4.76*	10.65**	-1.64	0.63	0.95	-1.11
	(1.80)	(2.46)	(4.51)	(3.19)	(1.49)	(3.37)	(2.97)
married	1.94	6.40**	-0.14	2.27	-0.91	3.82	-2.13
	(1.33)	(2.51)	(5.00)	(3.22)	(1.67)	(3.22)	(2.62)
self-employed	2.55	-5.79	7.87	-0.31	-1.11	-0.01	11.08
	(2.12)	(4.40)	(11.91)	(5.80)	(3.44)	(4.26)	(11.22)
not employed	-2.13	-3.12	6.74	-4.10*	0.05	-6.47*	-9.03***
	(2.19)	(2.56)	(4.68)	(2.32)	(1.90)	(3.57)	(3.13)
EQ-5D-5L [a]	-0.18**	-0.08	-0.23	-0.03	-0.06	-0.43***	-0.05
	(0.07)	(0.06)	(0.180)	(0.08)	(0.06)	(0.14)	(0.11)
awareness Q2	-1.09	1.36	-1.27	-2.83	-1.19	2.66	-1.15
	(0.56)	(3.28)	(5.96)	(3.78)	(1.83)	(4.42)	(3.83)
awareness Q3	-2.43	4.88	-4.50	-5.41	0.93	-2.87	-3.89
	(1.65)	(3.44)	(5.78)	(3.56)	(1.81)	(4.11)	(3.62)
awareness Q4	4.26	11.03**	2.04	-1.24	2.48	7.08	4.36
	(2.31)	(4.56)	(6.23)	(4.54)	(2.52)	(4.47)	(5.50)
no past exp.	-3.02	-3.21	-7.61**	-5.46*	2.25	-18.5***	-3.19
	(2.77)	(4.87)	(3.80)	(3.29)	(1.62)	(5.66)	(3.21)
HRAS Q2	-0.38	-1.01	-0.98	5.52	-0.66	-4.55	0.06
	(1.17)	(2.91)	(4.29)	(3.60)	(2.06)	(4.92)	(4.21)
HRAS Q3	-0.17	3.47	0.18	5.00	-0.83	-8.98*	-1.76
	(1.69)	(3.12)	(4.59)	(3.50)	(1.72)	(4.58)	(4.68)
HRAS Q4	4.92*	5.37	14.72**	5.94	1.55	1.91	-1.07
	(2.04)	(3.97)	(6.31)	(3.94)	(2.11)	(5.14)	(4.15)
constant	-10.8	-4.79	-34.07	-30.21	3.33	44.35	-13.55
	(10.7)	(26.10)	(30.5)	(27.89)	(12.45)	(29.98)	(30.63)
Observations	2,417	457	421	420	374	403	342
R^2	0.156	0.167	0.215	0.117	0.173	0.247	0.161
AIC	23,173	4,409	4,288	3,981	2,980	3,875	3,238
BIC	23,202	4,477	4,357	4,049	3,047	3,943	3,303
RMSE	29.263	29.553	38.644	27.120	12.716	29.031	26.859

Note: Q, quartile; Standard errors in parentheses; $* p < 0.10$, $** p < 0.05$, $*** p < 0.01$; Outliers defined as WTP > 5% of income; [a] sum score rescaled from 0 to 100.

Discussion

To estimate the value of an international integrated early warning system for infectious diseases and food-borne outbreaks aimed at increasing health safety, we developed a two-stage contingent valuation experiment. A survey containing the experiment was administered to balanced samples from Denmark, Germany, Hungary, Italy, the Netherlands, and the UK. The share of respondents indicating a WTP of zero varied between 7.3% in Italy and 23.2% in Hungary, of which most were protest zeros. Excluding protest answers and outliers (with a WTP exceeding 5% of income), the elicited overall mean monthly WTP per household was €21.80 (median=€10.00). This value ranged from €8.89 (median=€3.85) in Hungary to €28.33 (median=€13.42) in Denmark. The corresponding standard deviations were substantial, expressing either diverse or ill-formed preferences. Differences between countries can partly be explained by the variation in purchasing power, Hofstede's cultural dimensions and trust in public institutions. The results, in general, indicate that the majority of respondents see a certain value in the early warning system. Regression analyses showed that throughout countries and models, income, as expected, was the most important determinant of the WTP values elicited in our experiment.

Limitations and validity

Before discussing the implications of our findings, we must acknowledge several limitations inherent to our analysis and the contingent valuation approach. Individual WTP estimates are susceptible to the design and framing of a WTP exercise.[40] For instance, by instructing respondents to consider other similar contributions to inform their WTP (see online Appendix C), we may have introduced a possible anchor point for some individuals, biasing our results.[153] Furthermore, by listing very serious (but low probability) threats like Ebola, Sars and bird flu in the description of what the system aims to contain and mitigate, respondents may have overestimated the potential health gains of the system. However, as the aim of our analysis was to capture gains in their feelings of safety in the valuation, this is of less concern. A possibly more problematic concern of this type of contingent valuation studies is the respondent's sensitivity to the chosen payment scale.[40,153] It has also been reported that valuations are relatively insensitive to framing the payment as a monthly or yearly instalment.[154]

To reduce the effects of such potential biases in our study, we tested two additional versions of our survey, varying payment scale and frequency of payment, and chose the survey version, which provided the most internally consistent results. The two-stage approach, asking respondents for a value they would definitely pay and a value they would definitely not pay before the actual valuation, also aims to reduce midpoint bias and scale sensitivity. Including a "more" option in the payment scale was intended to decrease endpoint bias. A further limitation of WTP studies, in general, is the hypothetical nature of the experiment itself. Whether respondents would indeed pay the elicited amounts in real life is questionable. Research has shown that hypothetical WTP questions typically lead to an overestimation of actual WTP.[154,155]

A limitation specific to our analysis is that the actual unit of valuation, an international integrated early warning system for infectious diseases and food-borne outbreaks, is also a hypothetical construct, as it is not in existence yet. The survey included a concise description of its general purpose (online Appendix A), but we did not provide any more detailed information on the actual functioning and effectiveness of such a system. We also do not know about respondents' expectations concerning potential future (health safety) benefits through such a system. Besides these tangible benefits, individuals might also have incorporated potential improvements in the feeling of safety due to the system in their WTP valuation, as well as other benefits. Respondents may have unrealistic expectations regarding the potential (health) benefits of the early warning system, leading to distorted WTP valuations. However, as mentioned earlier, individuals make similar decisions without complete knowledge of real risks or benefits when deciding on specific types of insurance coverage. In both cases, they include perceived risks and benefits in their decision-making.

One further noteworthy limitation of our study is the exclusion of individuals aged 65 and above. One could argue that the WTP would be higher in the excluded group as they are in general more vulnerable to infectious diseases. We do not find strong evidence for this hypothesis, considering that the coefficients of age-squared are not significant in general and small in size. Future studies could investigate this age group further.

Despite these limitations, there are several aspects, which generate some confidence in the validity of the chosen design and our findings. For instance, the included warm-up exercise eliciting the WTP for the market good shoes provided plausible results, with means ranging between €61.09 in Hungary and €138.54 in Denmark. These results suggest that the respondents understood the question format and the WTP elicitation exercises and answering formats. Results from the survey module about contents insurance furthermore indicate, that the stated WTP, i.e., the perceived value of the system, somewhat corresponded to an actual WTP. This can be inferred from comparing elicited WTP and premiums paid for home contents insurance (as reported by respondents) in relation to respondents' indication of their relative value. For example, respondents who indicated the values of the warning system and home contents insurance to be similar, the mean difference in premiums and estimated WTP was €5.28 with half of the differences lying within an (admittedly arbitrary) €10 range.

The results from our regression analysis moreover demonstrated that, in general, WTP behaved as expected. WTP increased with income and to some extent with awareness of outbreaks and risk aversion. The positive effect of the level of trust in public institutions and the significance of the included cultural dimensions were further reassuring findings.

Considering that most of the mentioned limitations are inherent to willingness-to-pay approaches, one could wonder whether other methodologies, not based on stated preferences, would have been the more appropriate methodological choice. Such methods could entail using valuations of statistical life years or monetising the potential health gain using QALY threshold values. However, there are two main reasons, concerning feasibility (mainly due to the limited current knowledge about the warning system) and scope of the analysis, why this is not the case. First, the statistical life year approach requires the availability of certain types of (international) data, which, at this stage of the COMPARE project are not available, yet, if they can be provided at all, or are difficult to obtain in general. Using QALY thresholds, on the other hand, requires the availability of threshold values in all countries of interest, while explicit threshold values are only available for the UK and the Netherlands. Noteworthy in this context is also that some of the estimates of the value of statistical life years and QALYs are based on willingness-to-pay studies, which had similar drawbacks as our study. Second, and more importantly, applying either of these methodologies would shift the focus exclusively on valuing direct health gains of the warning system. We opted for the current methodology and operationalisation as we intended to also capture the society's valuation of the perceived feeling of safety that comes with the envisaged system. The applied methodology is admittedly not perfect, with results also reflecting beliefs and imperfect information of respondents, which, next to methodological limitations, warrants caution in their interpretation.

Implications of study findings

Notwithstanding this, our study provides results, which have implications for policymakers and stakeholders in the context of interventions increasing health safety of the population in European countries. For instance, in a more general sense, our results indicate that most European citizens seem to value an early warning system when using additional taxation as a payment mechanism in an experimental setting.

Aggregating our WTP estimates to a national or international level can inform discussions about appropriate funding of the warning system, given current knowledge and perceptions of the effectiveness of such a system. While we stress the explorative nature of our study, based on the median WTP estimates from Table 12, the relevant number of households (excluding the share of protesters), and assuming 50% of those households would be eligible to pay the additional tax, an aggregate WTP of €6.5bn for all six included countries per year would be estimated (see online Appendix F). Considering that health care spending on prevention is rather modest in the included countries (compared to spending on curative care), this may be considered a high amount. While the mentioned limitations related to estimates of individual WTP apply to the aggregate as well, it can help to give such a number a bit more context. A study from the Netherlands in 2007 estimated a yearly comprehensive national spending on preventive measures aimed at infectious diseases of €261.46 per capita (inflation-adjusted €333.16 in 2018).[156] This national spending includes vaccinations but in particular, infrastructure aimed at protection from infectious diseases like waste disposal and clean water technologies. An early warning system could be seen as an add-on to this existing infrastructure. On a per capita level (17.3m citizens), the aggregated WTP in the Netherlands would be €23.71, which corresponds to 7.1% of the previously calculated comprehensive national spending on infectious disease prevention.

Assuming that the calculated aggregate WTP corresponds to actual yearly costs and that the early warning system would reduce the burden of disease of influenza of 81.8 DALYs per 100,000 in the six included European countries by 20%, e.g., through rapid sequencing of new types of influenza and timely vaccinations, the costs per DALY averted would amount to €164,190. This ratio does not yet include DALYs averted in other infectious diseases or food-borne outbreaks, nor does it account for the economic burden of such outbreaks, which can be considerable,[136] or more intangible benefits like the increased feeling of safety.

In terms of the methodology used in this study and the specification of the contingent valuation approach, future research should investigate the use of more precise assumptions about the actual benefits of such a warning system and how this impacts the WTP valuations. As soon as information on the effectiveness of a COMPARE like warning system is available, it could also be of interest to explore the value of the direct health gains, e.g., using a statistical life year approach or different national estimates of the value of health gains.

Conclusion

Overall, our analysis provided first estimates of the perceived value of this type of early warning system in European countries. While the used approach is clearly not without limitations, the results of our analysis can be relevant to policymakers when discussing investments in health safety on a European level in general, and an early warning system for infectious diseases in particular. However, future research will have to provide further information on what this system would look like, the costs associated with installing and maintaining such a system, and how effective it would be at actually increasing health safety, i.e., reducing the risks of pandemics and outbreaks as well as mitigating their impact, among European citizens. Only then, it is possible to assess whether the investment in such a system is money well spent and health and welfare improving.

6

Did the COVID-19 pandemic change the willingness to pay for an early warning system for infectious diseases in Europe?

SFW Himmler, NJA van Exel, WBF Brouwer

European Journal of Health Economics. 2022; 23:81-94

Abstract*

The COVID-19 pandemic highlights the need for effective infectious disease outbreak prevention. This could entail installing an integrated, international early warning system, aiming to contain and mitigate infectious diseases outbreaks. The amount of resources governments should spend on such preventive measures can be informed by the value citizens attach to such a system. This was already recognized in 2018, when a contingent valuation willingness to pay (WTP) experiment was fielded, eliciting the WTP for such a system in six European countries. We replicated that experiment in the spring of 2020 to test whether and how WTP had changed during an actual pandemic (COVID-19), taking into account differences in infection rates and stringency of measures by government between countries. Overall, we found significant increases in WTP between the two time points, with mean WTP for an early warning system increasing by about 50% (median 30%), from around €20 to €30 per month. However, there were marked differences between countries and subpopulations, and changes were only partially explained by COVID-19 burden. We discuss possible explanations for and implication of our findings.

* The online version of this article (https://doi.org/10.1007/s10198-021-01353-6) includes supplementary materials relating to appendix tables A1 to A8 and figures A1 to A4 in this chapter. The code for the analysis presented in this chapter can be accessed through the Open Science Framework: https://osf.io/2wtve/.

Introduction

The current COVID-19 crisis and previous infectious disease outbreaks show that uncontrolled pandemics can have disastrous global consequences,[157,158] with recent estimates putting the global price tag of COVID-19 in terms of economic and disease consequences at 8 to 16 trillion dollar.[159] At the same time, the likelihood of the occurrence of pandemics, as well as the magnitude of their impact in terms of disease and economic burden, can be lowered drastically if appropriate measures are taken.[160] Pandemic prevention could, for example, consist of reducing the likelihood of zoonosis outbreaks themselves in different ways. It was estimated that a global strategy, involving measures like limiting deforestation and wildlife trade, as well as implementing early detection and control measures, would require yearly investments of over 20 billion dollars, but could be highly cost-effective.[159] Aiming to prevent and control zoonosis outbreaks early on, however, is only one, although important, piece of the puzzle of prevention of and preparedness for future pandemics.[161] Governments around the globe, independently, or on a supranational level, must ask themselves how to prepare for, or prevent, a next pandemic or similar health crisis. This also involves choices regarding how much funds can or should be invested in pandemic prevention measures, not knowing when and if such an event will occur again. As was pointed out by Chilton et al. (2020),[162] (welfare) economic tools can assist "in the process of building preparedness for similar future events". Next to calculations like those presented by Dobson et al. (2020),[159] information on society's willingness to pay for pandemic prevention measures can provide useful information in this context.

This was recognized also before the COVID-19 outbreak. Himmler et al. (2020)[163] attempted to estimate the willingness for improvements in health safety provided by an international, integrated early warning system for identifying, containing, and mitigating large infectious disease outbreaks. Using a willingness to pay (WTP) experiment with samples from six European countries, they found a mean monthly WTP of €21.80 (median €10.00) per household for such a system, with large differences across countries (from € 8.89 in Hungary to €27.32 in Italy). The data for this study was collected in March 2018, two years before the COVID-19 outbreak, using hypothetical scenarios.

The current COVID-19 crisis provided the opportunity to test whether this willingness to pay would change now that a pandemic is reality rather than only a hypothetical scenario. Hence, we replicated the study by Himmler and colleagues in the spring of 2020, at a time when COVID-19 cases were increasing exponentially, economic consequences of the pandemic became clearer, and strict governmental measures were already imposed across Europe. This replication entailed fielding the same survey, using the same sampling approach, and same procedures to estimate and analyze WTP, to ensure maximum comparability between the two studies.

While one might expect the perceived value of such a warning system for infectious diseases to increase during a pandemic, as its usefulness may be more apparent and individuals' preferences more informed, we aim to confirm this and explain any differences across the two time points by re-running the same models and comparing results. We also want to investigate whether differences are related to the impact of COVID-19, in terms of cases per 100,000 population, and the stringency of governmental measures at the time of sampling. In addition, by replicating different WTP scenarios in a new context (the pandemic), this study addresses common methodological questions regarding stated preference studies in general and contingent valuation WTP studies in particular, namely their sensitivity to scope and context.[164] This may provide further insights into the validity of estimates obtained through such studies and, hence, their policy relevance. While neither of the two experiments may necessarily elicit the "true" WTP, the unique set-up allows us to at least attempt a more nuanced interpretation of the WTP data, which ultimately may also inform public investments into pandemic prevention.

Methods

Survey and willingness to pay scenarios

In the spring of 2020, we re-fielded a survey including a willingness to pay experiment, which was initially administered in 2018 to samples from the UK, Denmark, Germany, Hungary, Italy, and the Netherlands.[163] The same online panel provider was used (Dynata) to obtain samples of 500 individuals from each of these countries (as in the 2018 survey). We aimed for the same number of respondents as in the 2018 survey to ease comparability of WTP estimates across the two data collections. Using quota sampling, the country samples were aimed to be representative in terms of age and gender for the working age population (aged 65 or younger). The 2020 survey additionally included a sample of 500 individuals from northern Italy (defined as the regions north of Lazio and Umbria), where COVID-19 cases, mortality and lockdown measures were most severe at the time of sampling.

The contingent valuation procedure consisted of a two-step payment scale approach followed by an open-ended question to elicit the maximum willingness to pay for an integrated, international early warning system for infectious diseases. The original survey consisted of eight scenarios specifying different levels of risk reduction and (health) consequences of an outbreak; respondents all completed two basic scenarios first and were randomly assigned two of the six remaining scenarios. The 2020 survey only included the four most realistic scenarios for the current context. The flow of the WTP scenarios in the 2020 survey and the corresponding per country target samples available for analysis across the two timepoints is shown in Figure 14. Each respondent completed three scenarios. All respondents completed the 'System' and 'Base case' scenarios first (names not shown to respondents) and were then randomized to either the 'Certainty' or the 'Death' scenario.

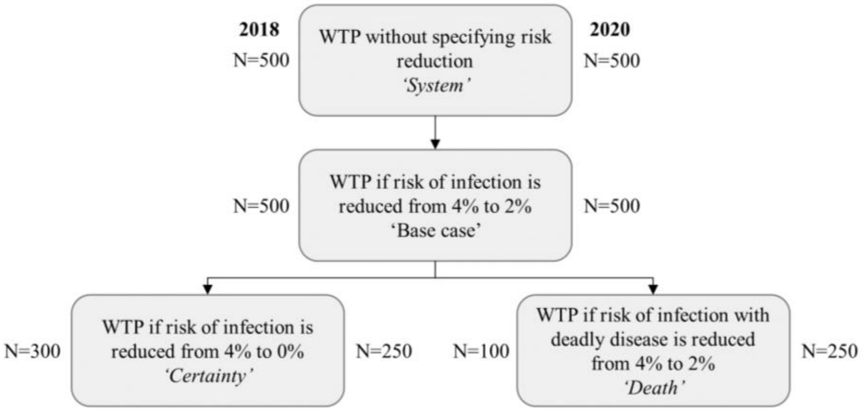

2018
N=500
WTP without specifying risk reduction
'System'
2020
N=500
N=500
WTP if risk of infection is reduced from 4% to 2%
'Base case'
N=500
N=300
WTP if risk of infection is reduced from 4% to 0%
'Certainty'
N=250
N=100
WTP if risk of infection with deadly disease is reduced from 4% to 2%
'Death'
N=250

Figure 14: Willingness-to-pay scenarios and target samples per country for 2018 and 2020 survey. The 2018 survey included additional scenarios not shown here. The target sample for Italy in 2020 was 1,000, 500 of which from northern Italy (north of Lazio and Umbria).

In the 'System' scenario, it was outlined to respondents that establishing and maintaining an international integrated warning system aimed at containing and mitigating infectious disease and food-borne outbreaks, like Ebola, SARS, bird flu and salmonella (COVID-19 was added to the 2020 survey), is costly. Respondents were then asked to assume that the funding would take place through national taxation via monthly instalments starting immediately and were then asked how much they would be willing to pay per month for having this international, integrated warning system. In the 'Base case' scenario, a 4% risk of becoming infected with a virus within the next three months was specified. If infected, health would reduce from a good to a bad health state for the duration of one year, which were described using EQ-5D-5L profiles corresponding to utility values of 0.887 and 0.574 (using the UK tariff from Devlin et al., 2018 for all countries).[32] Respondents were then asked to imagine that the risk to become infected can be reduced from 4% to 2% through the early warning system and subsequently had to state their willingness to pay analogous to the previous scenario. In the 'Certainty scenario', the risk reduction was specified to be from 4% to 0%. In the 'Death' scenario, the risk and the reduction were the same as in the 'Base case' scenario, but the consequence of an infection would be immediate death instead of a health deterioration for the duration of one year. Before each of the risk scenarios, respondents were made familiar with the concept of risk and probability using visual aids, similar to Bobinac et al. (2014).[166] More details about the structure of the survey, the design of the WTP exercise, and type of survey administration and data collection can be found in the preceding study.[163]

Timing of data collection

In addition to the available data of 3,140 observations from the 2018 survey, we were able to collect WTP responses from 3,979 individuals in March/April 2020 of whom 650 also participated in the 2018 survey. Figure 15 shows the timeline of the data collection in relation to the prevalence of COVID-19 cases and the timing of restrictive policy measures in each of the included countries.[167,168] Most of the sample was collected in the last weekend of March 2020. This was at a time when the number of cases was increasing rapidly in all included countries and restrictive policy measures, with a significant impact on peoples' lives and daily activities, had been in place for a couple of weeks, with Hungary as the exception for both. The prevalence of COVID-19 in that period was consistently two to three times higher in Italy compared to Germany, the Netherlands, the UK, and Denmark, which all experienced a similar trajectory. Throughout the sampling period, the confirmed COVID-19 cases remained at a low level in Hungary. These considerable differences between countries need to be kept in mind when interpreting the results of our analysis.

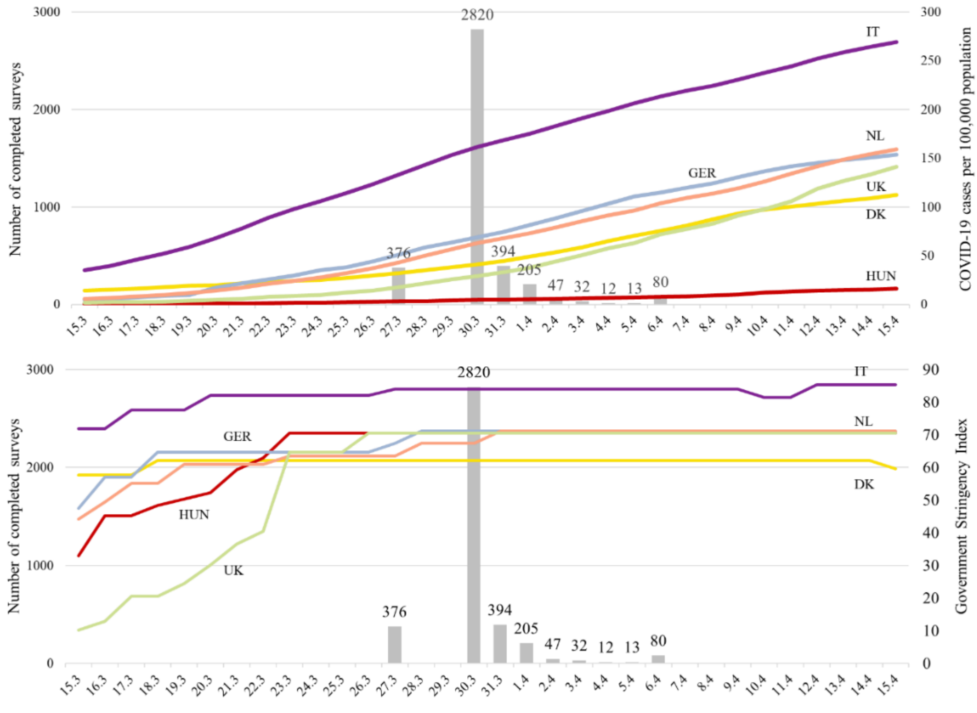


Figure 15: Timing of survey responses, COVID-19 cases, and Government Stringency Index of measures. Case data from ECDC (2020).[168] Stringency index from Oxford COVID-19 Government Response Tracker.[169]

Data analysis

Before analyzing the WTP data, several steps were undertaken to facilitate a valid comparison across countries and timepoints, taking the results presented in Himmler et al. (2020) as a reference. First, income and WTP values from the UK, Denmark and Hungary were converted to Euro values using the average exchange rates from March 2018 and 2020, respectively. Second, using the same criteria as in the previous study, protest answers (defined as zero response justified by warning system being a government task), and outliers (defined as a monthly WTP larger than 5% of monthly household income, which was deemed an unrealistic WTP) were identified in each of the four scenarios and excluded from the WTP analysis of the respective scenario. Third, using the country specific consumer price indices for March 2020 from Eurostat, 2020 income and WTP values were deflated to 2018 values.[170] Fourth, country specific monetary values were purchasing power adjusted using the latest available purchasing power parities from 2018 and the European 27 countries index as a base.[146]

All monetary values reported in this study therefore represent PPP adjusted values in 2018 prices. To facilitate the comparison of regression results across the two time points when pooling country level data together, we weighted the 2020 WTP scenario observations according to the country composition in 2018. Although all country samples at both time points initially consisted of roughly 500 respondents, this was necessary as the data cleaning (protest answers and outliers) lead to unbalanced samples at both time points. The additional sample of 500 respondents from Norther Italy was omitted from these pooled regressions on the representative samples.

After these steps, mean and median willingness to pay were calculated for each scenario for all countries, the repeated sample (of respondents who participated in 2018 and 2020), and the total sample. To check whether a change in WTP would be related to changes in WTP (or ability to pay) across all kinds of products, we calculated the difference in WTP between the two time points for a pair of shoes, the included warm-up WTP exercise. To facilitate a comparison, the shoe WTP values were rescaled to the mean WTP in the 'System' scenario in 2018.

To test if the changes in WTP across the timepoints were significantly different and not a result of differences in the samples between the two time points, we ran ordinary least squares (OLS) models in the following form, pooling the data from 2018 and 2020:

$$WTP_{isc}=\alpha_{sc}+\beta_{sc}*y2020+\gamma_{sc}*SES_{isc}+\varepsilon_{isc} \qquad (8)$$

Willingness to pay values for the four scenarios *s* and the samples *c* (countries and combined sample) were regressed on the year indicator *y2020*, controlling for the vector *SES* which contains the following variables: log of monthly household income, age, age-squared, gender, level of education, marital status, and employment status. Only the resulting *β* parameters, which indicate the change in WTP due to the COVID-19 outbreak, will be reported. Standard errors were clustered on country level for the combined sample regression. A similar fixed effects regression, excluding the time invariant covariates, was run for the sub-sample of individuals, which were observed at both time points to account for time invariant unobservables.

Himmler et al. (2020) conducted linear regression analysis to examine whether factors influencing WTP were in line with theoretical considerations, as well as previous empirical findings of WTP determinants. To test if there were meaningful shifts in the importance of these determinants between 2018 and 2020 samples, and whether these could be linked to the COVID-19 outbreak and its consequences, we repeated the analysis for both timepoints. WTP values from all four scenarios (Figure 144) and countries were combined, increasing the number of observations and therefor the statistical power to detect significant changes.

WTP was modelled as a function of the same vector *SES* as in equation (8); health status, as measured using the sum score of the EQ-5D-5L; the level of awareness of outbreaks; whether individuals or a family member have been exposed to an infectious disease outbreak before or not; and the health-risk attitude of respondents, which was assessed using the sum score of the six-item version of the health-risk attitude scale and included as quartile indicators.[171] The awareness variable, which was originally a sum score of 12 Likert-scale questions, was split into three sub-scores to provide more nuance: personal risk perception and behavior, societal consequences of outbreaks, and risk and response. For the full questions, see Appendix Figure A1.

It is important to note that the (statistical) comparison of regression coefficients from two independent samples is inherently difficult, even if the data generating process is the same and the samples should be comparable. Consistent inference on the parameters across the 2018 and 2020 samples was facilitated through Stata's 'suest' command.[172] This command provides estimates for seemingly unrelated regressions using a joint variance–covariance matrix of all parameters. This allowed us to compute t-tests comparing coefficient estimates across 2018 and 2020 samples. Standard errors in the regressions were clustered on individual level to account for the dependence of WTP responses within an individual.

A significance level of 10% was used throughout the analysis. The statistical analysis of the data was performed using Stata 16.0 (Stata Corp. 2019. Stata Statistical Software: Release 16. College Station, TX: Stata Corp LP).

Results

Sample characteristics

Sample characteristics and country composition are presented in Table 14. While observations were otherwise equally distributed across countries, we obtained a larger sample for Italy in 2020, with 394 respondents specifically from north Italy. Information on the response rate and completion rate was not provided by the sampling agency. There were no considerable changes in overall respondent characteristics between the two sampling periods except an increase of the share of dependent employed individuals from 54% to 58% and an increase in monthly household income by 5.4% (after adjusting for inflation and purchasing power). Important to note is that the level of income was lower in the 2020 sample for Hungary (Appendix Table A1 contains country level means). The sub-sample of individuals, who were observed at both timepoints had a significant lower level of income and was significantly older than the full samples in 2018 and 2020. Furthermore, the repeated sample has been previously exposed to infectious diseases to a lesser degree (Appendix Table A2). These differences imply that the individuals who participated twice in the survey, represent a specific selection of individuals. In this hereafter called ‘repeated sample’, respondents from the UK, Germany and Italy are furthermore overrepresented, as less individuals who already participated in 2018 could be sampled from Denmark and the Netherlands. Appendix Table A4 shows the dataset conditioning for the different parts of the analysis.

Table 14: Characteristics of full sample and repeated sub-sample across timepoints.

	Full sample		Repeated sample	
	2018	2020	2018	2020
Monthly income in €[a]	2,917 (3,765)	3,052 (4,969)	2,571* (2,038)	2,726* (4,564)
Age	42.2 (14.0)	42.7 (13.1)	43.8* (12.2)	45.8* (12.1)
Female	0.51	0.51	0.50	0.50
No finished sec. education	0.03	0.03	0.03	0.02*
Finished high school	0.57	0.57	0.58	0.58
Tertiary education	0.40	0.40	0.39	0.40
Married	0.58	0.57	0.56	0.59
Employed	0.54	0.58	0.58*	0.60
Self-employed	0.10	0.11	0.12*	0.11
Unemployed	0.06	0.08	0.08*	0.07
Homemaker	0.07	0.06	0.07	0.07
Student	0.10	0.06	0.05*	0.04*
Retired	0.09	0.08	0.07*	0.08
Unable to work	0.04	0.04	0.03	0.03
Country				
UK	0.18	0.16	0.19	0.19
DK	0.16	0.13	0.10	0.10
GER	0.17	0.16	0.19	0.19
HUN	0.16	0.13	0.16	0.16
IT	0.17	0.17[b1]	0.25	0.15[b1]
		0.10[b2]		0.10[b2]
NL	0.17	0.16	0.11	0.11
Observations	3,140	3,979	650	650

Note: [a]In 2018 PPP. Income information was available for 2,772 and 3,608 respondents in full sample and 578 and 584 respondents in repeated sample. [b1]South and [b2]North Italy. * $p < 0.10$ in independent t-tests comparing repeated to full sample in the respective year.

Changes in awareness, exposure, health-risk attitude, health, and well-being

To aid in interpreting the WTP results, we will first summarize some descriptive evidence on changes in contextual factors like awareness of outbreaks, health risk attitude, past exposure, and health and well-being between the March 2018 and March 2020 samples. More detailed descriptions of these factors and/or the corresponding results are provided in Appendix 1.

Overall, the awareness or perceptions of risks and consequences of infectious disease outbreaks increased (Figure A1). People feel more at risk compared to others, would be more willing to take precautionary measures advised by authorities, are more concerned about infectious diseases compared to other diseases, and are informing themselves about outbreaks more often. They are more aware of the damage such outbreaks can have on health, social life, and the economy, while agreeing to a much higher degree that outbreaks are a major public health concern (65% to 81%). Interestingly, the share of individuals, who think that the risk of outbreaks cannot be lowered by taking precautionary measures, remained almost the same (7% to 6%). A striking observation is that even during the COVID-19 pandemic, 45% of respondents agreed with the statement that outbreaks originate in other countries, and it would be their responsibility to deal with them (44% in 2018), dismissing the need for an international response.

The sample of March 2020 was, in general, slightly more health-risk averse with a mean HRAS score (range 6-42) of 30.1 (SD 5.8) compared to the 2018 sample (28.8, SD 5.7). This shift can largely be explained by respondents agreeing to a greater extend with the statement "To enjoy good health now and in the future, I am prepared to forego a lot of things" (49% to 62%) (Figure A2). The relative increases in awareness and health risk aversion between 2018 and 2020 were similar across all countries (Appendix Table A3). The highest levels thereof were observed for Italy for both time points.

The share of individuals reporting that they themselves or a family member have been exposed to an emerging infectious disease or foodborne outbreak in the past decreased from 19% to 16% in the total sample. In Italy, this share increased from 13% to 16% and 18% in north and south Italy, respectively. The large differences in self-reported exposure between countries (Appendix Table A3) may partly be a result of a different interpretation of the question (in 2018 the share varied from 10% in Denmark to 62% in Hungary). Similarly, observed decreases in the rate between the two timepoints could reflect more accurate responses in 2020, as respondents were likely more knowledgeable about the subject area due to the COVID-19 outbreak. In terms of the impact the COVID-19 outbreak on self-reported health, life satisfaction and capability well-being, we did not observe any meaningful changes between the 2018 and 2020 sample (Appendix Figure A3).

Willingness to pay across countries and timepoints

Of the total of 20,606 WTP values across the four scenarios, 1,104 were classified as outliers, and 1,643 as protest answers (Appendix Table A5 provides scenario level information). Dropping these observations lead to a WTP analysis sample of 17,859 observations. The share of protest answers and zero WTP responses were in general lower in 2020 compared to the 2018 sample, apart from the 'Death' scenario. The largest drop in the share of protest answers was observed for Hungary (e.g., from 17% to 7% for the 'System' scenario. The share of WTP values classified as outliers on the other hand, increased for almost all scenarios and countries.

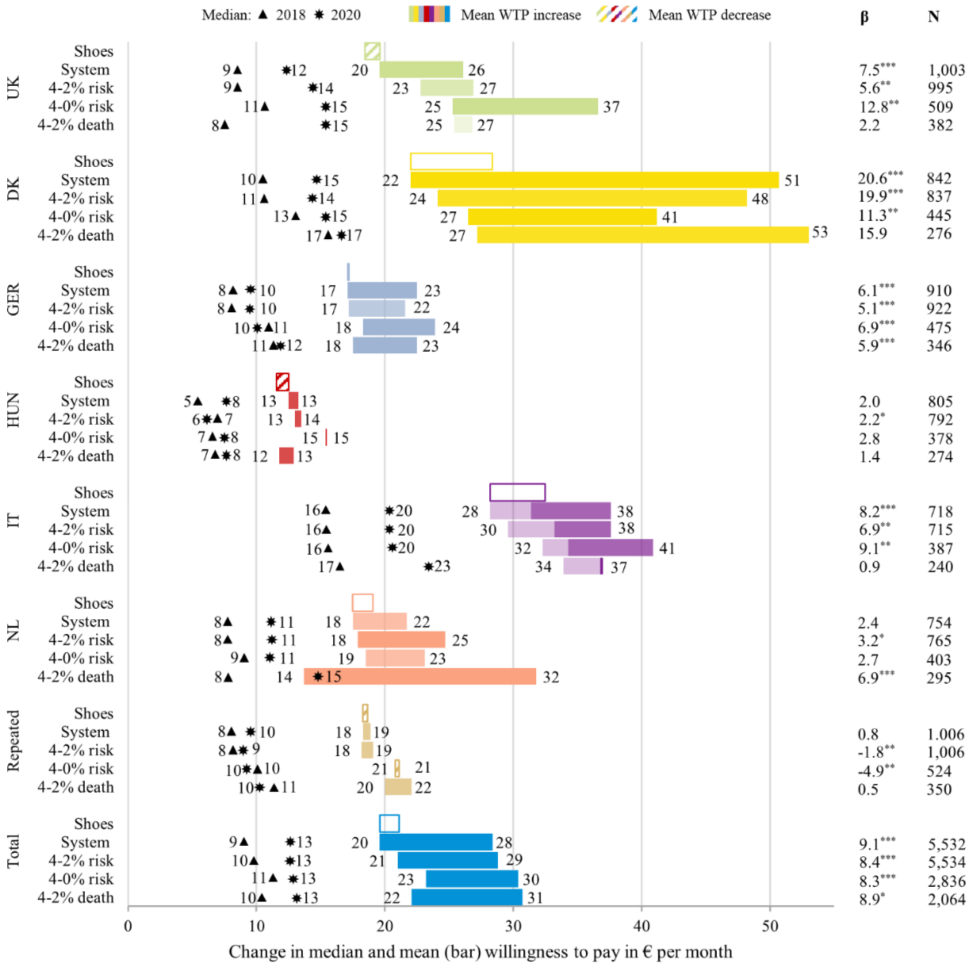


Figure 16: Changes in willingness to pay for an early warning system across scenarios, countries and timepoints. WTP in 2018 PPP. Changes in mean WTP from 2018 to 2020 represented as bars. WTP for shoes as reference and rescaled to 'System' 2018 values. Deep colour bars for Italy represent additional WTP in northern Italy compared to southern Italy and 2018. Total sample values weighted to maintain same country composition in aggregate. β parameters represent coefficients of the y2020 dummy variable from regression on the pooled sample, controlling for log of income, age, gender, education, and marital and employment status (Equation 8). N is the number of observations in the respective regressions. * p < 0.10 ** p < 0.05 *** p < 0.01.

Figure 16 presents mean and median WTP for an early warning system for infectious diseases across scenarios and countries, comparing 2018 values to the values obtained during the COVID-19 outbreak in 2020. The Figure also includes the scenario and country level estimates of β, the timepoint dummy from the pooled regression analysis (equation 8). There is large variation in WTP values across countries, scenarios and timepoints. Important to note is that the country specific WTP values were rather stable across the four scenarios, despite the differences imposed in the scenario description (Figure 14). The total mean WTP increased by between 30 and 40%, depending on the scenario, corresponding to an additional monthly contribution of 7€ to 9€ (baselines values were 20€, 21€, 23€ and €22 for the four scenarios). The total median monthly WTP increased by between €1.6 to €3.6 (15 to 40% increase). The total variation in elicited WTP values more than doubled in each of the four scenarios. In 2018, the variation in WTP in the 'System' scenario was 28.6, while in 2020 the standard deviation was 71.2.

The largest increases across all scenarios were found for Denmark. There, WTP in the 'System' scenario almost doubled, even after accounting for differences in socio-economic characteristics (baseline 2018 value of €22, β-coefficient €20.6). Besides for the 'Certainty' scenario in the UK, moderate increases in monthly WTP of up to €10 were found in the remaining countries. The WTP was lowest in Hungary, with values remaining almost stable across timepoints (maximum monthly WTP increase of €2.8 and not significant). There was a larger increase in monthly WTP in northern Italy (up to €9.1) compared to the south, with the Italian sample having reported the highest levels of WTP in 2018. Interestingly, WTP was stable or even decreased in the subset of observations, which were observed twice. Results from the reference point included, WTP for a pair of shoes (rescaled to mean of the 'System' results for 2018), indicated that willingness and ability to pay, in general, slightly increased, except for Hungary and the repeated sample, across the two timepoints.

As COVID-19 cases and governmental measures increased over the period of data collection (Figure 15), whether certain sub-samples were collected particularly early on or later may have impacted WTP. However, we found no worrisome pattern in our data.

Figure 17 plots willingness to pay values for the 2020 sample against the country aggregate number of COVID-19 cases and the government stringency index. There seems to be a positive relationship between number of cases and the WTP for an early warning system. Interestingly, the occurrence of extreme values seems to decrease over time (higher number of cases equals later timepoint as number of cases was consistently increasing over the sampling period) within most countries. A positive relationship was also observed for WTP and the government stringency index, although the variation in the strictness of government measures was much smaller (Figure 17).

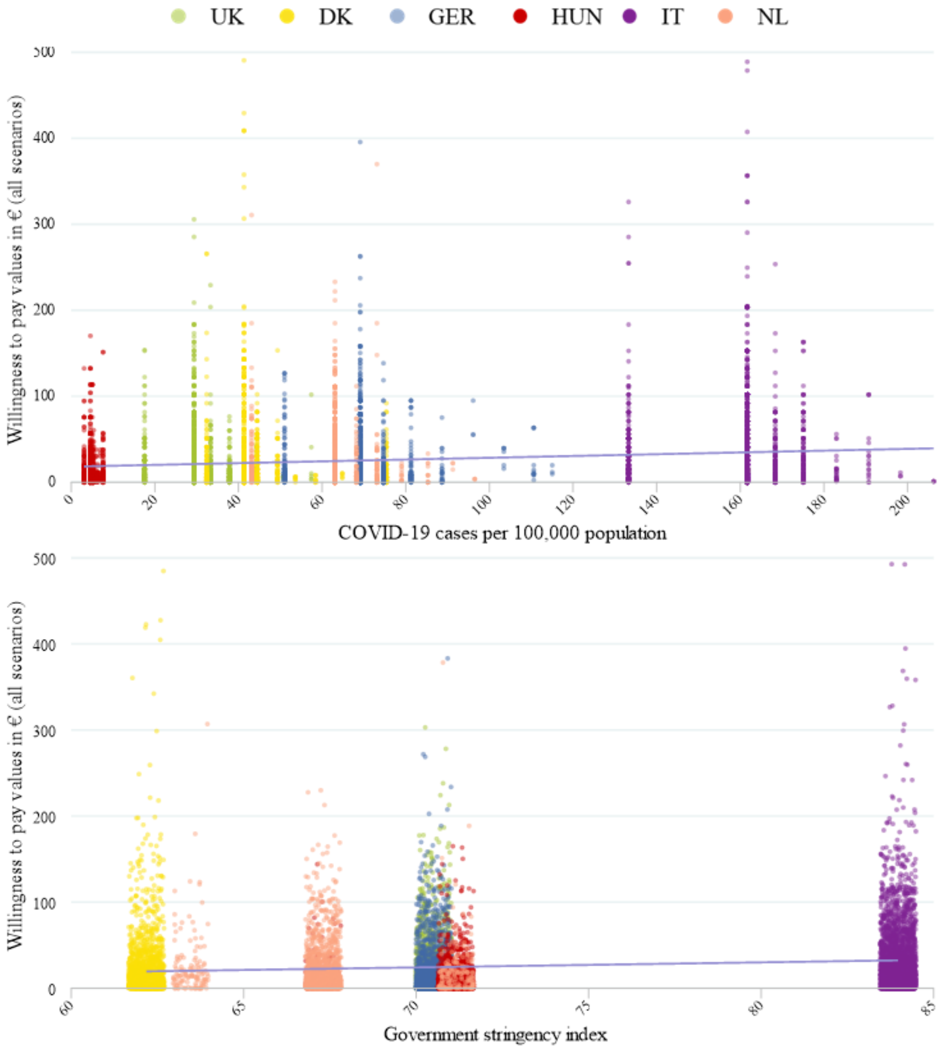


Figure 17: Willingness to pay during COVID-19 outbreak in relation to number of cases and measures. Case data from ECDC (2020).[168] Stringency index from Oxford COVID-19 Government Response Tracker.[169] Horizontal line represents linear fit. Random variation added to GSI (jitter).

Determinants of willingness to pay

Table 15 presents results of the seemingly unrelated estimation procedure, allowing the comparison of coefficients of WTP determinants across 2018 and 2020 sample. A structural difference in the overall associations was also confirmed by a Chow-test, which rejected the null hypothesis of equality of coefficients in the 2018 and 2020 samples (chi-squared: 56.37, $P < 0.01$). The included variables explained a larger share of the variance in WTP in the 2020 regression. While the directions of associations with WTP remained stable for some variables (log-income, being female, tertiary educated, or self-employed, and past exposure), the estimated coefficients switched sign for variables like being married or unemployed. Besides these changes, large and significant differences in coefficient size were found for log-income, self-employment, being in the highest health-risk aversion quartile. Personal risk perception and behavior also played a larger role for the 2020 WTP values. That coefficient estimates generally increased may partly be explained by the larger WTP values and the larger variation in WTP values found in 2020 compared to 2018 (standard deviations in all four scenarios doubled).

Appendix Table A6 presents the results for the subsample of repeated observations. As differences in WTP between the two timepoints were considerably less pronounced in this sample, changes in the importance of determinants occurred less frequently. The Chow test further confirms no structural change in overall coefficients between the two timepoints (chi-squared: 19.21, $P = 0.57$). In general, the variables followed a similar pattern compared to the full sample. A notable exception is that the coefficient of self-employment did not increase. No structural change in coefficients estimates was also found in the subsamples of respondents from Italy based on the Chow test (chi-squared: 22.53, $P = 0.13$).[†] Interestingly, the coefficient of past exposure decreased (Appendix Table A7).

[†] Excluding the additional sample from northern Italy in 2020, as we did not have comparable data for this sample for 2018.

Table 15: Determinants of willingness to pay across time points

	2018		2020		*P-value*
Socio-economics status					
Log income	9.41***	(1.01)	33.57***	(3.70)	< 0.001
Age (Δ5 years)	-5.38***	(1.39)	-2.87	(2.71)	0.399
Age-squared	0.18**	(0.08)	-0.03	(0.15)	0.220
Female	-3.55***	(1.03)	-2.78*	(1.67)	0.682
Tertiary education	1.21	(1.10)	4.23***	(1.48)	0.099
Married	2.17**	(1.03)	-15.87***	(3.12)	< 0.001
Self-employed	2.06	(2.09)	38.95***	(7.65)	< 0.001
Not employed	-2.25**	(1.14)	8.37***	(2.19)	< 0.001
EQ-5D-5L sum score (Δ5 points)	-0.89***	(0.22)	-0.01	(0.31)	0.019
Awareness of outbreaks					
Personal risk perception (Δ5 points)	6.02***	(0.82)	13.16***	(1.95)	0.007
Societal consequences (Δ5 points)	-2.14***	(0.80)	-1.52	(2.12)	0.784
Risk and response (Δ5 points)	-2.07	(1.29)	-17.31***	(3.90)	< 0.001
Past exposure	4.20***	(1.28)	2.37	(1.95)	0.432
Health risk attitude					
HRAS-SF Q2	0.12	(1.33)	-2.28	(2.40)	0.382
HRAS-SF Q3	-0.40	(1.33)	-1.27	(2.12)	0.729
HRAS-SF Q4	4.88***	(1.56)	12.18***	(2.46)	0.011
Observations	6,611		8,442		
Adjusted R-squared	0.190		0.278		
Chow test statistics	56.37***	*P < 0.01*			

Note: WTP values from all four scenarios as dependent variable. Standard errors were clustered on individual level and are presented in parentheses. Northern Italy subsample from 2020 excluded. Country dummies and constant omitted from table. Regression is weighted by 2018 country sample sizes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Discussion

Summary of WTP results

During the onset of the COVID-19 crisis in Europe, we repeated an experiment from 2018 by Himmler et al., which elicited the WTP for improvements in health safety provided by an international, integrated early warning system for identifying, containing, and mitigating large infectious disease outbreaks. Overall, we found statistically significant increases in mean monthly WTP by about 50%, depending on the specified WTP scenario (e.g., from €20 to €28 in the ‘System’ scenario), while the corresponding medians increased by about 30% (e.g., from €9 to €13 in the 'System' scenario). Differences between countries were more pronounced compared to the 2018 data collection. The largest increases in WTP were observed for the UK, Denmark, and Italy. We furthermore found rather stable WTP values in a sub-sample of individuals before and during the COVID-19 outbreak. Most of these individuals did not change, or only slightly, their WTP between the two timepoints (Figure A4).

Possible explanations of changes and patterns in WTP

The observed *moderate* increases in WTP for an early warning system for infectious diseases elicited pre-pandemic and during the first wave of COVID-19 in Europe may be interpreted in different ways. An optimistic interpretation is that the experiments set out to elicit WTP twice for the same good: an early warning system. The fact that the resulting WTP estimates at both points in time were not considerably different could signal that the anticipated risks and consequences of pandemics influencing the WTP during the first experiment were similar to the more informed ones during the second experiment. The higher awareness of outbreaks, and the risk and consequences of their occurrence, which we observed, then lead to respondents forming reasonable and realistic increases in WTP given their ability to pay. The elicited WTP estimates in 2020 then constitute an upper bound, as the COVID-19 pandemic and its consequences likely and hopefully remain an extreme variant of an infectious disease outbreak.

A more pessimistic interpretation would be that the chosen approach does not invite respondents to reveal changing preferences, for instance due to insensitivity to scale and scope in the elicitation technique. The methods used in, as well as the framing and scope of the experiment, may then not adequately reflect changes in 'actual WTP' following the COVID-19 outbreak. Although our data does suggest that there is some plausible sensitivity in our results and also some patterns that represent logical deviations from the initial WTP estimates, we cannot fully disentangle or refute the optimistic and pessimistic interpretations of our findings.

In terms of the differences in changes between countries, the finding that WTP in Italy increased more than for instance in Germany, the Netherlands, and Hungary, may not be unexpected given that Italy was hit hardest by the pandemic in March/April 2020. We did not find considerable changes in WTP in Hungary, which could be related to the fact that, at the time of data collection, it was the least affected country. On an individual level, WTP values during the COVID-19 outbreak seem to be determined to a higher extent by respondent characteristics. This relates to the potential impact of such a pandemic on individuals' lives and livelihoods, or the perceived individual (health) risks, as well as attitudes towards these risks. The most notable change in WTP determinants was observed for being self-employed. This is in line with first evidence from Germany, indicating that self-employed individuals were hit hardest by the pandemic in terms of economic consequences.[173] Gross monthly income was reduced for 59% of self-employed (vs. 15% of employed), with a median reduction of €1,500.

At the same time, we also found patterns that may be considered more unexpected. For instance, the increase in WTP in Northern Italy was small in relation to the severity of the crisis there during data collection. This contradicts the explanation that WTP is importantly influenced by the severity of the crisis. A potential explanation for this finding could be that Italians in this region were relatively dissatisfied with the COVID-19 response of their government, as well as with the assistance from the international community.[174] This could have decreased their trust in the possibility of an effective integrated international early warning system. It is also important to note that WTP was already highest in Italy in the 2018 sample, arguably leaving less room for further increases. Likewise, finding a high WTP for the early warning system in Denmark does not appear to correlate with the COVID-19 burden in that country (Figure 15). There, it may relate both to higher incomes and the high level of trust in national public institutions and the government,[152] which also prevailed during (the early phase of) the pandemic.[175] A further possible explanation for country-level WTP changes not being directly related, or at times being even reversely related to the burden of COVID-19, is that contributing to a preventive system now, actually does not help to overcome the *current* crisis. Respondents may feel that the current crisis should be given priority in terms of public expenditures, especially if the COVID-19 burden is severe. In that sense, it is good to highlight the difference between our study considering *preventative actions,* compared to curative, or mitigating actions, for instance asking about the WTP for a vaccine. Again, given the setup of our study we cannot be conclusive regarding these potential influences.

Another aspect, which may have influenced WTP values elicited during the pandemic, could be that individuals anticipated an economic downturn, and the personal consequences thereof, as a result of the pandemic. Therefore, they might be less willing (or able) to pay additional taxation. However, results from the non-health-related reference point included in our survey (WTP for a pair of shoes) and the income information indicated that, on average, the ability to pay (for everyday products at least) was not yet significantly affected by the COVID-19 outbreak. Indeed, the first noticeable economic consequences of the pandemic likely occurred after our sampling period in March 2020. Also, respondents in the second data collection may have been more aware of the fact that such a system would help to avoid later losses in income. This could have resulted in an increased willingness to pay, since they were more aware of the benefits of such a system for their own economic situation.

These explanations may also have caused WTP values to be fairly stable in the subgroup of respondents who completed the survey at both moments in time. In addition, it is important to note that respondents in this subgroup had a lower income (Table 14) and had a lower level of previous exposure to infectious diseases (Table A2). The country composition in this sample also did not reflect the original sampling quotas (equal across countries) and the individuals, who were observed twice, were different compared to the samples in their respective countries (Table A2). That we found a small *decrease* in ability to pay (as measured via the WTP for shoes scenario) for the repeated sample, which is in contrast to what was found for most included country samples, further highlights that this sample represents a specific selection of individuals.

Additional findings

We found notable shares of protest answers and zero responses. Moreover, a large share of respondents at the time were not convinced of the need of an international response during the COVID-19 outbreak. Furthermore, a significant proportion of respondents were still not aware of (or ignored) the seriousness of the societal impact of an outbreak, as well as the fact that precautionary measures could decrease the risk of outbreaks (see beginning of results section). These individuals may therefore disapprove of the governmental measures taken and might be hesitant to take up vaccination if available.[175,176]

Finding no differences in well-being and life satisfaction between the 2018 and 2020 samples may be somewhat surprising. The fact that the survey was fielded at a time when the *full* impact of the crisis on individuals' well-being and the economy at large was not clear to respondents (Figure 15) may help to explain this. A study from Germany, comparing individuals from a large panel sample across April 2020 and April 2019, also did not find a change in life satisfaction due to the COVID-19 outbreak.[177] Capability well-being, as assessed by the ICECAP-A, which specifically aims to measure capabilities and opportunities, was also not lower in our second survey compared to the first, even though the COVID-19 related lockdowns imposed quite drastic limitations on individuals' freedom and rights. It is interesting to see how such outcomes will evolve during the crisis, especially when these restrictions are imposed for longer periods of time.

Limitations of the analysis

Similar limitations as were outlined in more detail in the first study[163] apply to the current study as well. These relate to more general limitations of stated preferences and contingent valuation approaches, such as hypothetical response bias, insensitivity to scope, and framing effects.[164,178] These limitations are particularly important when the good under valuation is less tangible to respondents. This clearly applies here, as the early warning system for infectious diseases and its consequences are still hypothetical. Respondents therefore may have had difficulties in imagining such a system and its potential costs and benefits. Insensitivity to scope, which has been shown to exist before in the health domain using a similar set up,[41] was evident in our analysis considering the WTP results for the different presented scenarios. The small difference across scenarios may also be a result of respondents anchoring their WTP on their valuation of the first presented WTP scenario ('System' scenario). Insensitivity to scope may also explain the relative insensitivity of observed WTP values to the changes in circumstances over time, i.e., the COVID-19 outbreak. Regarding the hypothetical nature of the experiments, the following is important to note: the COVID-19 outbreak made the pandemic scenario more *real*. However, whether that made the presented

WTP scenarios (Figure 14) more *realistic* for an average respondent, is unclear. If scenarios were not recognized as relevant for the COVID-19 situation (e.g., because not so many people will be infected or die, or because the health states were not deemed plausible in relation to COVID-19), the scenarios possibly remained as hypothetical as in the first data collection.

In terms of the comparison of the 2018 and 2020 WTP values, it needs to be acknowledged that ideally, we would have resampled the full 2018 survey population. This would have enabled us to compare the same individuals within representative country samples. Although attempted this turned out not to be possible, and hence we needed to assume that the representative samples from 2018 and 2020 did not differ too much in terms of unobserved characteristics, which would have influenced WTP. For example, the data collection during the COVID-19 outbreak on such a topic could have attracted specific populations who would sooner select into participating in a survey on this topic. On the other hand, we took several steps, to enable a (valid) comparison, like PPP adjusting, accounting for inflation, weighting sample compositions or controlling for observable characteristics in the year-dummy regressions.

A final limitation concerning our sample is that we do not have WTP information from individuals aged 65 and older, which are the ones with the highest risk of serious health consequences due to a COVID-19 infection. The sample of 65 and younger may be seen as primarily (though clearly not exclusively) affected by economic consequences. This may be a reason why we found that age and health were not significant determinants of WTP in 2020, while self-employment and unemployment were. One might hypothesize that the largest changes in WTP over time may have occurred in the risk group of individuals aged 65 and above, which were not included in our samples. This reduces the generalizability of our findings, especially in contexts where the financing of an early warning system would be based on contributions from all citizens, including those older than 65 years of age.

While not a limitation, it is important to note that our study focused on European countries and similar experiments may have led to very different WTP results in other parts of the world even after PPP adjustment. Knowledge about COVID-19 and the public's perception of the pandemic and the associated risks, factors likely influencing WTP, vary widely across the globe.[179–181] In addition, the measures taken against COVID-19 between for example Europe and East Asia are different, thus may also translate to differences in WTP for an early warning system, as individuals would value efforts for either adopting or avoiding these measures more depending on individual values. One prominent example relate to the type of isolation used (or mandated) for mild COVID-19 patients,[182,183] where East Asian countries such as China adopted facility-based isolation with financial support and mental health counselling and lowered patients' anxiety to transmit virus to family members, yet may not be valued in Western countries such as the UK due to privacy infringements.

Conclusions

Repeating a European survey from 2018 eliciting the WTP for an early warning system aimed to prevent or mitigate outbreaks of infectious diseases, we found a higher WTP in 2020 as compared to 2018 in all countries except Hungary. We also observed a considerable increase in the heterogeneity in elicited values (both within and between country samples). Respondents showed some sensitivity to scope and to the context of the experiment (the COVID-19 outbreak), oftentimes in expected directions. However, the sensitivity to scope and context varied and should be interpreted with caution (see e.g., Bobinac et al. (2012)).[41] Our results should therefore be taken to represent a range of WTP values rather than a precise estimate of some 'true' WTP for an early warning system. We also stress that the contingent valuation WTP method has notable limitations, especially given the abstract nature of an early warning system.

Nonetheless, also in the absence of clearly better alternatives, our study aims to provide a relevant indication of the societal valuation by European citizens of such an early warning system for infectious diseases. Conducting a back-of-the-envelope calculation aggregating the median WTP values at the country level (similar to the preceding study[163]),[‡] the *moderate* increases in individual WTP translate into sizable increases at a societal level. The implied yearly maximum 'willingness to be taxed' increased from €1.3bn to €1.9bn in the UK, or from €6.0bn to €7.8bn summed over all six countries (Appendix Table A8).

Which of these two estimates, ex-ante or during the pandemic, is considered more informative depends also on the expected context of future outbreaks, which likely will have less extreme trajectories compared to COVID-19. Moreover, our study was conducted in the early stages of the pandemic when both duration and full societal impact were still unclear. Together with further related research during the subsequent stages of the pandemic, and also information on the (cost-)effectiveness of measures to prevent and control infectious disease outbreaks, this may inform policy makers on the type and magnitude of possible investments to prevent future outbreaks or mitigate their consequences.

[‡] Multiplying the median yearly WTP in the 'System scenario' with the number of households per country, excluding the share of households with protest zero answers, and assuming that 50% of the remaining households are eligible for this form of taxation.

7

Estimating the monetary value of health and capability well-being applying the well-being valuation approach

SFW Himmler, NJA van Exel, WBF Brouwer

European Journal of Health Economics. 2020; 21:1235-1244

Abstract*

Background: Quality of life measures going beyond health, like the ICECAP-A, are gaining importance in health technology assessment. The assessment of the monetary value of gains in this broader quality of life is needed to use these measurements in a cost-effectiveness framework.

Methods: We applied the well-being valuation approach to calculate a first monetary value for capability well-being in comparison to health, derived by ICECAP-A and EQ-5D-5L, respectively. Data from an online survey administered in February 2018 to a representative sample of UK citizens aged 18–65 was used (N = 1512). To overcome the endogeneity of income, we applied an instrumental variable regression. Several alternative model specifications were calculated to test the robustness of the results.

Results: The base case empirical estimate for the implied monetary value of a year in full capability well-being was £66,597. The estimate of the monetary value of a QALY, obtained from the same sample and using the same methodology amounted to £30,786, which compares well to previous estimates from the willingness to pay literature. Throughout the conducted robustness checks, the value of capability well-being was found to be between 1.7 and 2.6 times larger than the value of health.

Conclusion: While the applied approach is not without limitations, the generated insights, especially concerning the relative magnitude of valuations, may be useful for decision-makers having to decide based on economic evaluations using the ICECAP-A measure or, to a lesser extent, other (capability) well-being outcome measures.

* The online version of this article (https://doi.org/10.1007/s10198-020-01231-7) contains supplementary material relating to Appendix A in this chapter.

Introduction

Health economic evaluations are increasingly used in health care decision making. In countries like the UK and the Netherlands, specifically cost-utility analysis is a frequently applied tool to inform the allocation of scarce (health care) resources, with the aim of optimising population health.[46] In recent years it has been questioned whether health, measured for example with instruments such as EQ-5D, is the appropriate maximand in all contexts of health care delivery. Sometimes, the benefits of care interventions may not be limited to health alone, and the aim of interventions may not be to restore or improve health, but rather to maintain or increase the well-being of patients.[21,22] The question of what we want to maximise appears especially relevant in the palliative and elderly care sectors, and in mental health and integrated social care.[19,20] The interventions in those areas may range from pharmaceutical interventions to home care and, in the context of multi-morbidity, combinations of treatments.

As a consequence, several instruments have been put forward, aiming to measure quality of life in a broader sense, which could be applied to broaden the evaluative space of health economic evaluations.[81] In this context, some researchers focused on an operationalisation of Amartya Sen's capability approach,[184] which emphasises the importance of individuals' ability to reach certain well-being states (capability) instead of being in these states (functioning). A prominent example is the ICEpop CAPability measure for adults (ICECAP-A), an instrument developed for assessing the capability well-being of the general adult population. The ICECAP-A measures capabilities in five dimensions with four levels each: (i) stability (ii) attachment (iii) autonomy (iv) achievement, and (v) enjoyment.[24] The measure was validated and tested in different contexts with promising results and continues to be validated further.[50–52,185–187] Moreover, it was shown that the ICECAP-A measures a broader construct and also comprises complementary information compared to common generic health utility measures like EQ-5D-3L and EQ-5D-5L.[188,189]

In the new Dutch pharmaco-economic guidelines specific attention is paid to broader outcome measures, in particular the ICECAP instruments.[190] This may not only increase their use in the context of economic evaluations of pharmaceutical and other interventions, but also brings up the issue as to how the results of such broader economic evaluations should be used in decision making. Indeed, the current (applications of) capability measures still raise important questions,[191] including how results from economic evaluations using capabilities, likely in the form of incremental cost-effectiveness ratios (ICERs), should be interpreted. Valuable in this context would be information on an appropriate threshold value for capabilities, analogous to the quality-adjusted life-year (QALY) threshold for health gains. While the monetary value of a QALY has been extensively studied, primarily using willingness to pay,[9,38] research on the monetary value of capability well-being is still lacking.

This study aims to fill this gap, by estimating a first monetary value of a year in full capability well-being, using the well-being valuation method to ICECAP-A index values in a representative sample of UK citizens aged 18 to 65. Using the same approach and sample, we furthermore provide estimates of the same kind for the monetary value of a QALY based on EQ-5D-5L data, facilitating a first comparison of the societal valuations of these constructs.

Methods

Conceptual model

The well-being valuation approach uses observational data to assess the experienced average impact of a change in a good on individuals' overall utility *u*, proxied by subjective well-being (SWB) or life satisfaction, and calculating the change in income necessary to maintain the same level of utility.[192] This obtained monetary valuation is also known as compensating surplus (CS). This regression-based approach circumvents the inherent drawbacks of willingness to pay experiments by not directly asking individuals for a monetary value of a certain good.[193,194] Applying the well-being valuation approach for estimating monetary values of capability well-being and health requires the following assumption about the relationship between health, capability and SWB: Individual's overall utility *u*, as proxied by SWB, is a function of health or capability well-being *Q*. Imposing this type of relationship on capabilities is in conflict with the normative position that capabilities go not only beyond health but also beyond utility and SWB.[195] While we do acknowledge that there is some evidence based on individual-level data in favour of this competing interpretation,[196] this is a necessary assumption due to the mechanics of the well-being valuation approach.

$$u(Q,Y,X)=SWB(Q,Y,X) \tag{9}$$

Utility *u* is furthermore determined by income *Y,* and certain individual and socioeconomic characteristics summarised in vector *X.* We followed a three-stage well-being valuation procedure, as previously formulated.[192,197] The three steps include separately estimating the impact of income and the good to be valued on SWB (steps 1 and 2) and then calculating the compensating surplus (CS) according to equation (10) (step 3):

$$CS=Y^0-e^{\left[\ln\left(Y^0\right)-\frac{Q'}{Y'}\right]} \tag{10}$$

Y' and *Q'* are the marginal effects of changes in income and health or capability on SWB, and Y^0 represents a representative level of population income.

Data and model specification

The data for the analysis originated from a cross-sectional survey of UK citizens, which was not specifically designed for this analysis and is, therefore, limited to individuals aged 18 to 65. Random sampling and survey administration were conducted by Survey Sampling International in February 2018 using an online survey format. The sample was aimed to be representative regarding age, gender and level of education and consisted of 1,512 individuals. The survey included inter alia questions about health, well-being, income, employment and marital status, religiosity and information about the health risk attitude of respondents (in the listed order).[198]

The impact of health *H* and capability well-being *CW* on SWB were estimated separately, due to their substantial overlap and likely collinearity. While it has been discussed before that estimating the effect of health on SWB is prone to issues of endogeneity,[79,199] it was not possible to address this issue adequately due to the limitations of the used data. Applying a previously used instrument for health – average health per socioeconomic cell – was not feasible, possibly a result of the small sample size,[200] SWB_i was assessed using Cantril's ladder, a one-dimensional life satisfaction instrument asking respondents to rate their life from worst possible to best possible life on a 0-10 scale.[69] The impact of health and capability well-being were estimated using ordinary least squares, assuming cardinality in the responses:[65]

$$SWB_i=\beta_0+\beta_1 H_i+\beta_2 \ln(Y_i)+\beta_3 X_i+\varepsilon_i \quad (11)$$

$$SWB_i=\alpha_0+\alpha_1 CW_i+\alpha_2 \ln(Y_i)+\alpha_3 X_i+\mu_i \quad (12)$$

Health of respondents H_i was measured via EQ-5D-5L utilities, applying the English EQ-5D-5L tariff estimated by Devlin et al. (2018).[32] Capability well-being, CW_i, was assessed via ICECAP-A index values.[24,47] Estimates for income Y_i were obtained by asking respondents to place their combined monthly household income before taxes into 12 prespecified intervals. In a follow-up question, respondents were asked to indicate exact amount within these intervals. Missing exact income amounts were imputed based on the sample means of the income interval selected in the first step, if applicable. X_i contains age, gender, education, marital status, and employment status, which have been shown to influence SWB.[201] Following further guidance from the literature, we also controlled for religiosity, measured by asking for the importance of religion on a 7-point Likert-scale, and religious affiliation.[202] Information on the health risk attitude of individuals[198] was included to partly account for personality.[203]

Income coefficient estimates in SWB regressions are likely endogenous due to reverse causality,[204,205] measurement error or omitted variables like working hours, or time spent away from family.[206] Instrumental variable (IV) approaches have been used to overcome this problem.[44,207] We, therefore, applied a two-stage least squares (2SLS) approach,[208] testing different available candidate instruments. In the final analysis, we used whether a household currently holds home contents insurance (*CI*) as an instrument for income *Y*. The logarithmic transformation of income was used to account for its diminishing marginal return on SWB.[209] The 2SLS approach took the following form:

$$SWB_i=\gamma_0+\gamma_1 H_i+\gamma_2 \ln(Y_i)+\gamma_3 X_i+\omega_i \quad (13)$$

$$\ln(Y_i)=\delta_0+\delta_1 CI_i+\delta_2 X_i+v_i \quad (14)$$

To be a suitable instrument, *CI* must be sufficiently correlated with income. Possible channels could be that purchasing the insurance is more affordable if income is higher, or that higher income could lead to the household containing more valuable objects, which increases the likelihood of obtaining *CI*.

The instrument should furthermore only be correlated with SWB through income. However, this is generally not testable.[208] It is unlikely that the presence of contents insurance (directly) influences individuals' SWB. The insurance effect of increased (financial) stability could be a possible channel. However, we found only a small and negative correlation between *CI* and the stability dimension of the ICECAP-A (r=-0.15). Maintaining *CI* could relate to personality traits like risk aversion, which might influence SWB. Nevertheless, we were directly controlling for risk attitude, which is furthermore merely weakly correlated with *CI* (r=0.14). Additionally, the obtained SWB values might not originate from the same individual, who decided about purchasing *CI*. Unfortunately, we had no information available to investigate this. Finally, *CI* could be indicative of possessing more valuable items or living in a nicer home, which does impact SWB.[201] However, we argue that these aspects are also, at least partly, mediated through income.

Coefficient estimates from equations (11) to (14) were used to calculate the compensating surplus (CS) for one QALY and one year in full capability well-being (YFC) according to the following equations:

$$\mathrm{CS(QALY)}=\frac{1}{\Delta \mathrm{H}}*\left[\mathrm{Y}^0\text{-}\mathrm{e}^{\left[\ln\left(\mathrm{Y}^0\right)\text{-}\frac{\beta_1}{\gamma_2}*\Delta \mathrm{H}\right]}\right] \tag{15}$$

$$\mathrm{CS(YFC)}=\frac{1}{\Delta \mathrm{CW}}*\left[\mathrm{Y}^0\text{-}\mathrm{e}^{\left[\ln\left(\mathrm{Y}^0\right)\text{-}\frac{\alpha_1}{\gamma_2}*\Delta \mathrm{CW}\right]}\right] \tag{16}$$

Where Y^0 was set to the sample's median yearly household income of £27,000, while ΔH and ΔCW represented incremental changes in health and capability well-being. It was necessary to impose incremental changes of *H* and *CW* since under the framework laid out in equation (10) the CS would be constrained at the pre-specified level of income.[197] The incremental approach mirrors contingent valuation studies, where willingness to pay for small health changes are aggregated to a full QALY.[38] The size of the incremental change Δ was set to 0.1, corresponding to half a standard deviation, which was found to be a reasonable approximation of the minimally important clinical difference for health-related quality of life measurements.[210]

Descriptive and regression analyses were performed using STATA 15.0 (Stata Corp. 2018. Stata Statistical Software: Release 15. College Station, TX: Stata Corp LP). 2SLS estimates were obtained using the ivreg2 package.[211] All monetary amounts presented in the following correspond to 2018 prices.

Robustness checks

The robustness of the estimates was examined testing the following specifications: First, to gain insights into the relevance of accounting for the endogeneity of income, the non-instrumented, standard OLS income estimate was used instead of the IV income estimate. Second, an income coefficient estimate from a study based on much richer data was used. We linearly rescaled the dependent variable from a 0 to 10 to a 1 to 7 interval to match the SWB measure used in the analysis by Fujiwara,[197], and applied his log-income coefficient estimate, as it was based on (random) lottery wins. Third, SWB was assessed via the multidimensional Satisfaction with Life Scale (SWLS) instead of Cantril's ladder,[70] with SWLS scores rescaled from 0 to 10 to facilitate comparison of coefficients. Fourth, the unweighted average of Cantril's ladder and SWLS on a 0-10 scale were used as a compound SWB measure, as it was previously suggested that such a compound measure could be more robust than either of the measures on its own.[72] Fifth, instead of using the weighted population tariffs for scoring EQ-5D-5L and ICECAP-A values, we used the unweighted and rescaled (0-1) sum scores of these measures to test the sensitivity of the estimates to applying population tariffs, as both tariffs were based on different valuation methods. In the sixth robustness check, the mapped EQ-5D-3L value set was used instead of the EQ-5D-5L value set, since the methodology applied for the latter has come under scrutiny.[212] In the seventh specification, Y^0 was set to the mean yearly income of £37,843, instead of the median income of £27,000. In the last two robustness checks, ΔH and ΔCW were set to 0.05 and 0.20, as the size of the increment may still be considered somewhat arbitrary.

Table 16: Characteristics of analysis sample and IV-sample

	Total sample		IV-sample	
	Mean	SD	Mean	SD
Cantril's ladder	6.4	2.0	6.9	1.8
ICECAP-A	0.75	0.20	0.79	0.176
EQ-5D-5L	0.84	0.21	0.85	0.205
HH income in £	37,843	56,729	45,200	78,838
Age	42.6	13.9	47.2	12.4
Female	51.8%		48.8%	
Tertiary education	45.4%		50.0%	
Marital status				
Married	59.5%		66.8%	
Divorced/widowed	9.2%		10.9%	
Never married	31.3%		22.3%	
Employment status				
Employed	54.8%		61.9%	
Self-employed	9.5%		9.2%	
Unemployed	5.5%		2.2%	
Homemaker	9.7%		6.1%	
Student	5.2%		1.0%	
Retired	9.5%		14.8%	
Unable to work	5.8%		4.9%	
Religious affiliation				
Christian	42.1%		49.8%	
Atheist	32.8%		29.5%	
Agnostic	13.0%		11.9%	
Muslim	3.8%		1.8%	
Other religion	8.4%		7.0%	
Importance of religion	2.8	2.0	2.8	2.1
HRAS	29.0	5.8	30.1	5.4
N	1,373		1,373	

Note: IV, instrumental variable; HH, household; Importance of religion measured on a 1 (low) to 7 (high) scale; HRAS, Health Risk Attitude Scale ranging from 6 (risk loving) to 42 (risk averse).

Results

Estimates for income, health, and capability well-being

After excluding 139 observations with no income information, and imputing income interval sample means for 358 respondents who only reported their income interval, the analysis sample included 1,373 individuals. There were no missing values in the remaining variables. This sample was comparable to the UK population aged 18 to 65 concerning most characteristics (Table 16). The reported average yearly gross income of £37,843 in the sample is lower than the UK average of £45,773 in 2018. The average

ICECAP-A index is slightly lower than previously observed in a general population sample, which included individuals above 65 with generally lower capabilities.[213]

Table 17: Results of OLS and IV regressions

	(I)		(II)		(III)	
	Health		Capability		Income-IV	
Log yearly income	0.495***	(0.065)	0.308***	(0.054)	2.201***	(0.638)
EQ-5D-5L	2.665***	(0.305)			2.310***	(0.378)
ICECAP-A			6.234***	(0.243)		
Age	-0.026	(0.029)	-0.006	(0.024)	-0.004	(0.037)
Age-squared	0.0003	(0.000)	0.0001	(0.000)	0.0001	(0.000)
Male	-0.011	(0.093)	-0.012	(0.075)	-0.068	(0.119)
Tertiary education	0.038	(0.094)	-0.085	(0.076)	-0.395*	(0.199)
Divorced or widowed	-0.358*	(0.168)	0.078	(0.132)	0.256	(0.304)
Never married	-0.536***	(0.121)	-0.033	(0.096)	0.202	(0.306)
Self-employed	0.100	(0.180)	0.117	(0.139)	0.451	(0.249)
Unemployed	-0.579*	(0.231)	-0.275	(0.190)	0.661	(0.546)
Homemaker	-0.257	(0.169)	-0.028	(0.133)	0.387	(0.308)
Student	-0.357	(0.247)	-0.589**	(0.226)	-0.084	(0.365)
Retired	0.537**	(0.188)	0.115	(0.148)	0.864***	(0.253)
Unable to work	-0.514	(0.277)	-0.541**	(0.197)	0.672	(0.534)
Atheist	0.245	(0.138)	0.182	(0.111)	0.268	(0.168)
Agnostic	0.097	(0.161)	0.095	(0.139)	0.038	(0.202)
Muslim	-0.462	(0.303)	0.013	(0.241)	-0.330	(0.307)
Other religion	-0.013	(0.172)	0.101	(0.145)	-0.095	(0.235)
Importance of religion	0.147***	(0.031)	0.097***	(0.025)	0.148***	(0.038)
HRAS	0.077***	(0.009)	0.033***	(0.007)	0.0687***	(0.011)
Constant	-2.858**	(0.954)	-2.657***	(0.767)	-20.60**	(6.687)
N	1,373		1,373		1,373	
Root MSE	1.662		1.345		2.021	
R-squared	0.334		0.564		-	
Kleibergen-Paap rk LM					21.55***	
Kleibergen-Paap rk Wald F					21.63***	
Test for endogeneity					10.65***	

Note: HRAS, Health Risk Attitude Scale; Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Coefficients from the separate health and capability regressions as described in equations (11) and (12) are shown in columns (I) and (II) of Table 17. Parameters estimates for EQ-5D-5L, and ICECAP-A were positive and significant, (2.665 and

6.234), meaning that health and capability have the expected positive impact on SWB. The signs of the coefficients of most control variables corresponded to findings from the literature.[44,201] Coefficient estimates from the 2SLS IV regression are shown in column (III). Around a third of respondents (N=516) reported that their household holds contents insurance. The log-income coefficient was 2.201. Control variables deviated slightly between the models, namely in a higher positive impact of being retired, a negative impact of education and no effect of marital status and unemployment.

Kleibergen-Paap rk Wald F statistic (21.832 with Stock-Yogo critical 10% value 16.38) and Kleibergen-Paap rk LM statistic of (21.746, p<0.001), indicated that the used instrument was not weak or under-identified. This was further substantiated by a significant coefficient (p<0.001) of *CI* in the first stage regression (Appendix A). The characteristics of the IV sample were reasonably similar to the full sample (Table 16), with slightly higher levels of life satisfaction, capability well-being and income. Testing for the endogeneity of log income revealed that the variable should not have been treated as exogenous (p<0.001).

Implied monetary values and results from robustness checks

The resulting monetary valuations of one QALY and one YFC were £30,786 and £66,597, respectively. The relative size of the monetary value of capability well-being compared to health was thereby estimated to be 2.2. Coefficients estimates and the corresponding monetary valuations for the conducted robustness checks are shown in Table 18. First, not instrumenting for income led to considerably larger monetary estimates of one QALY (£112,336) and one YFC (£193,305). Second, applying the income coefficient from Fujiwara (2013), who used lottery wins, led to slightly higher monetary estimates compared to the base case. Third, using SWLS instead of Cantril's ladder provided an almost identical monetary value for one YFC, while the value of one QALY was reduced to £20,988. Fourth, the use of the compound SWB score averaged out differences in coefficients and monetary valuations between the use of Cantril's ladder and SWLS as SWB proxies. Fifth, employing sum scores of EQ-5D-5L and ICECAP-A resulted in slightly higher estimates of the value of one QALY and conversely, slightly lower estimates for one YFC. Applying the mapped EQ-5D-3L tariff reduced the monetary valuation of one QALY to £25,487. In the last three robustness tests, the income model had to be recalculated. As in the base case, the instrument passed under- and weak identification tests. Seventh, replacing median income by mean income increased the valuations to £43,149 and £93,343, respectively. Altering the imposed incremental change of 0.1 index points to 0.05 reduced the monetary estimates slightly while imposing a 0.2 incremental change led to higher estimates compared to the base case. Throughout model alterations, the monetary equivalent value of one YFC exceeded that of one QALY by a factor of around two, with the robustness check utilising SWLS as SWB proxy as an outlier.

Table 18: Base case monetary estimates and robustness to alternative specifications

	Base case	No IV	IV Fujiwara[a]	SWLS[b]	Comp. SWB[c]	Sum scores[d]	EQ5D-3L tariff	Mean income	Increm. 0.05	Increm. 0.20
Coefficients										
Log income	2.201	0.495	1.103	2.633	2.417	2.255	2.197			
EQ5D-5L	2.665	2.665	1.599	2.131	2.398	2.858	2.179			
ICECAP-A	6.234	6.234	3.740	7.488	6.861	5.881	6.234			
Value in £										
1 QALY	30,786	112,336	36,431	20,988	25,498	32,141	25,487	43,149	31,717	29,031
1 YFC	66,597	193,305	77,651	66,828	66,723	61,979	66,597	93,343	71,305	58,384
Rel. size	2.2	1.7	2.1	3.2	2.6	1.9	2.6	2.2	2.2	2.0

Note: All reported coefficients significant on the 1% level; [a] Income coefficient from Fujiwara et al. (2013), other coefficients from rerunning regressions with rescaled SWB; [b] Rescaled from 0 to 10, instrument passes under- and weak identification test; [c] Unweighted average of Cantril' s ladder and SWLS as SWB proxy, instrument passes under- and weak identification test; [d] EQ-5D-5L and ICECAP-A sum scores scaled from 0 to 1.

Discussion

Findings and related literature

Applying the well-being valuation method, we obtained a first estimate of the monetary value of ICECAP-A-derived capability well-being for the UK. We furthermore calculated the monetary value of health and were able to compare the valuations of one QALY and one YFC directly. The empirical challenge inherent to the chosen approach is the endogeneity of income, which we tried to overcome using whether a household holds contents insurance as an instrument for income. In the base case model specification, this yielded monetary valuations of £30,786 for one QALY and £66,597 for one YFC, corresponding to a ratio of 2.2. The conducted robustness checks produced relative magnitudes of these monetary valuations ranging from 1.7 to 2.6.

The calculated monetary value of a QALY lies within the range of estimates from the international willingness to pay literature, which on aggregate produced a trimmed mean and median estimate of £63,777 and £20,834 (in 2010 pounds).[38] UK specific estimates from Mason et al. (2009) and Baker et al. (2010) ranged from £24,219 to £70,896 and £16,000 and £24,805 (in 2010 pounds), respectively.[42,214] In the only other application of the well-being valuation method for this purpose to date, the monetary value of one QALY in Australia was estimated to be A$42,250 (£20,797) and A$67,022 (£32,990) for short and long-term health gains using 2015 prices.[44] The relative size of the reported monetary value of well-being (A$112,000 or £55,130) compared to one QALY was 1.7, not dissimilar to what we observed in our analysis.

Limitations

Although our results appear to have some face validity and are reasonably robust to model specifications, we need to acknowledge several limitations. On a more conceptual level, the chosen approach relies on the assumption that SWB is an appropriate proxy for individuals' utility. This may be a strong assumption, as SWB (or happiness) is not the only thing that people care about and preferences outside of SWB maximisation exist.[192] Nevertheless, based on the findings from subjective well-being research, as for example summarised by Diener et al. (2018),[215] we argue that SWB matters enough to be able to use it as a proxy for welfare. At the same time, we must acknowledge that the validity and reliability of SWB measures have been questioned before. These concerns were addressed in detail for example by Veenhoven (2012).[216] What we can infer from our own analysis, is that the choice of SWB instrument does have an impact on the monetary estimates (Table 18), although observed differences were not substantial. The SWLS appears to capture a different part of SWB than Cantril's ladder does. Differing results are likely a consequence of the SWLS containing two questions, which are more related to the past ("So far I have gotten the important things I want in life" and "If I could live my life over, I would change almost nothing"), while Cantril's ladder only asks about SWB at present, which is more consistent with the present based well-being valuation approach.[70] The well-being valuation literature so far does not provide guidance on the appropriateness of one- or multi-dimensional

SWB measures, or the use of a composite of both. This should be examined in future research.

A further limitation is that we had to deviate from the intended three-stage well-being valuation approach in two ways:[192,197] First, including control variables in order to prevent omitted variable bias conflicts with the idea of using total causal effects in calculating the monetary valuations as outlined before.[192] In the analysis by Fujiwara (2013), the difference in unemployment coefficients between a model without any covariates and a model controlling for several variables was minimal (-0.441 and - 0.436). Removing all control variables from models (I) to (III) generated monetary estimates for one QALY and one YFC of £33,914 and £63,156, respectively, close to the base case estimates. Second, and potentially more problematic, we assumed exogeneity of both health and capability well-being due to the lack of suitable instruments. When health was instrumented in a previous analysis, the estimated impact of a change in health decreased slightly.[200] Assuming this would also hold in our context, our monetary valuations represent overestimations.

It is furthermore inherently difficult to demonstrate that the used income instrument (contents insurance) satisfies the exclusion restriction assumption. In the second robustness check, we employed the log-income coefficient of Fujiwara (2013) for the UK, as an external reference point, after basing the analysis on the same SWB scale.[197] While not without limitations, his estimate, based on large scale panel data and exploiting random income shocks like lottery wins, can be considered as close to causal estimates as it gets when using non-experimental data. The reported log-income coefficient of 1.103 is comparable to the estimate we obtained when repeating the analysis on the same SWB scale of 1.321. Monetary estimates increased by around 20% (Table 18). Judging from this comparison, it appears that our instrument performs reasonably well.

The extent to which our results are generalisable to the general UK population is unclear, as our sample did not include individuals aged 65 and above. Previous research suggests that functional limitations and social functioning, which are more related to the ICECAP-A, could be more relevant to the elderly than typical health dimensions, like morbidities or pain.[217] To test this, we included an interaction term for the respective quality of life index and age to the base case models. We observed a positive and significant coefficient of 0.031 ($p=0.042$) for an interaction term between ICECAP-A and age, while the interaction coefficient of EQ-5D-5L and age of 0.021 was not significant ($p=0.355$). This indicates that omitting the elderly may have introduced a downward bias for the value of one YFC in comparison to the value of one QALY. Furthermore, due to relying on data from online survey panels, the individuals in the sample, in general, were quite healthy, with an average EQ-5D-5L index of 0.837 (SD 0.21). We do not know how the lack of sufficient observations at the lower end of the scale influenced our overall results. Lastly, we lacked information on the household size of respondents, which precluded the use of equivalised household income, to facilitate the comparability across household compositions.[44,206]

Interpretation and implications of the results

While the calculated values for one QALY and one YFC varied across the conducted robustness checks, their ratio fluctuated at around two. As well-being measures were designed to capture quality of life beyond health, it is explicable that the monetary value of well-being in general lies above the value of health alone. That this also holds for capability well-being could have been expected but had not yet been confirmed before (at the same time, it may be more difficult to achieve similarly sized increases on these broader measures). This information is relevant in the context of interpreting results of economic evaluations using broader outcome measures, which may be relevant in a range of interventions (from pharmaceuticals to palliative care) that have benefits not fully captured in conventional QALY measures.

The interpretation of the relative magnitude of the monetary estimates of one QALY and one YFC deserve further attention, considering that the EQ-5D-5L and the ICECAP-A are anchored on two different scales. The former is anchored on a 0 to 1, dead to full health scale, with the possibility of health states below zero.[32] The latter ranges from 0 to 1 for no capability to full capability, where death implies no capabilities, but no capabilities, in turn, does not necessarily imply death.[33,47] While it is plausible that on the higher end of the scale, capabilities go beyond health on an underlying overall quality of life continuum, it is less clear on the lower end of the scale, as having no capabilities could be equivalent to death, but also lower or higher in terms of overall quality of life. This may have implications for the comparability of the monetary valuations, as the imposed incremental change in health and capability of 0.1 may represent either a larger or smaller difference in the underlying utility. Future research could investigate these issues further, for instance, by focusing on the behaviour of SWB scores at very low levels of capabilities and health.

If capability well-being, as measured by the ICECAP-A, is included in future economic evaluations in areas where a focus on health is potentially too restrictive to capture all relevant benefits of an intervention, the here presented results could give a first indication about a cost-effectiveness threshold. In practice, ICERs calculated using ICECAP-A index values could be compared to the here estimated monetary value of a YFC. Our estimates are especially relevant for countries that relate their threshold to the societal monetary value of health or wellbeing gains, like the Netherlands.[9] In other countries, like the UK, thresholds are conceptually more related to the marginal cost-effectiveness of current spending.[218] Conceptually, this limits the direct applicability of our results in the UK, while it is noteworthy that obtaining opportunity cost based monetary estimates for capability well-being seems to be a challenging task.

Future research should aim for confirming our findings for the absolute and relative monetary valuation of capability well-being in general, either by employing alternative approaches, like willingness to pay or discrete choice experiments or by applying the well-being valuation method to other, preferably richer data sets. Prerequisite for the latter should be the availability of potential instruments for income. On a different note, while there are first applications, more conceptual and theoretical work is needed about whether, when and how capability well-being should be included in health economic evaluations.[111] One open question for example is, whether full capability or a sufficient level of capability, which was established recently, should be considered as the objective of interventions.[219] Nevertheless, and to conclude, the results of our analysis may be useful as a first estimate of a threshold value for a YFC that can be used when making decisions based on economic evaluations using the ICECAP-A, or to a lesser extent, other (capability) well-being outcome measures.

8

The Value of Health - Empirical issues when estimating the monetary value of a QALY based on well-being data

SFW Himmler*, J Stöckel*, NJA van Exel, WBF Brouwer
* Both authors contributed equally to this study

Health Economics. 2021; 30(8): 1849-1870

Abstract[*]

Decisions on interventions or policy alternatives affecting health can be informed by economic evaluations, like cost-benefit or cost-utility analyses. In this context, there is a need for valid estimates of the monetary equivalent value of health (gains), which are often expressed in e per quality-adjusted life years (QALYs). Obtaining such estimates remains methodologically challenging, with a recent addition to the health economists' toolbox, which is based on well-being data: The well-being valuation approach. Using general population panel data from Germany, we put this approach to the test by investigating several empirical and conceptual challenges, such as the appropriate functional specification of income utility, the choice of health utility tariffs, or the health state dependence of consumption utility. Depending on specification, the bulk of estimated € per QALY values ranged from €20,000-60,000, with certain specifications leading to more considerable deviations, underlining persistent practical challenges when applying the well-being valuation methodology to health and QALYs. Based on our findings, we formulate recommendations for future research and applications.

[*] The online version of this article (https://doi.org/10.1002/hec.4279) contains supplementary material relating to appendices A1 to A4. The code for the analysis presented in this chapter can be accessed through the Open Science Framework: https://osf.io/b8nsz/.

Introduction

During the ongoing COVID-19 pandemic, many citizens for the first time directly observe scarcity of goods in the health care sector in terms of testing, ventilation, vaccination capacity, and the prioritisation of services under binding capacity constraints. This scarcity and the broader societal consequences of the pandemic has revealed many difficult trade-offs between health and the economy, and between the needs of different patient groups within the health care sector. While the current attention to such matters is unprecedented, policy makers are confronted with many of these trade-offs also in non-pandemic times. To make informed decisions on policy options, however, requires decision makers to weigh up health and economic consequences, aiming to ensure maximum benefit or minimal harm. Welfare economic tools like cost-benefit-analysis can aid decision makers in this process by providing relevant and clear information to openly address the nature of the trade-offs being made.[162,220,221]

Cost-benefit analyses entail measuring and valuing gains and losses (benefits and costs) in monetary units, thereby allowing a holistic perspective on societal trade-offs and identifying which policy option is socially most preferred. In the context of interventions and policies affecting population health (though not necessarily aimed primarily at health), cost-benefit analysis therefore requires to obtain estimates on the monetary equivalent value of health, from here onwards denoted as v_{Q}.[222]

In the narrow health care context, v_Q, depending on the jurisdiction,[223] constitutes an important parameter in health technology assessment. There, value for money considerations are often operationalised using cost-utility analysis, where a new technology's costs are compared to its expected health gain, measured using Quality Adjusted Life Years (QALYs).[224] Equation (17) formulates a generalisation of the corresponding decision rule, with ΔQ denoting the health gain (in QALYs) and Δc_t the total costs compared to the alternative treatment:

$$\frac{\Delta c_t}{\Delta Q} < v_Q \qquad (17)$$

This cost-effectiveness ratio (ICER) is acceptable, if it lies below v_Q corresponding to one QALY (In the broader cost-benefit framework, this QALY equivalent v_Q value can be used for transforming health gains into monetary benefits).[9] While the use and empirical foundation of such threshold values within health care vary across jurisdictions,[34,225] estimating the level of v_Q corresponding to one QALY, also for the purpose of cost-benefit analysis, is challenging and has been attempted using various methods (see background section). In this endeavour, Huang, Frijters, Dalziel, and Clarke (2018) were the first to conceptualise and apply the well-being valuation approach for estimating a QALY equivalent v_Q, providing estimates of A$42,000 (€28,000) to A$67,000 (€45,000).[44] This method is based on the marginal rate of substitution between income and health. Further exploration of the approach is needed to be able to judge whether the corresponding estimates are indeed helpful for informing v_Q. This paper aims to make the following contributions: Firstly, by applying a similar approach as Huang et al. (2018) and using data from a different context, we generate further insights regarding the validity and reliability of the well-being valuation method for determining v_Q. Secondly, we aim to address some empirical and methodological challenges associated with applying the well-being valuation method in general and for valuing QALYs in particular, which were not fully addressed in previous studies. By using German data, an additional contribution lies in providing information on v_Q for a context in which such estimates are scarce, a result of German health authorities not (explicitly) basing their reimbursement decisions on the framework outlined in equation (17). Instead, the trade-off between Δc_t and ΔQ is discussed and determined in closed-door price negotiations between health authorities and the manufacturer. The methodological uncertainty around estimating v_Q has been cited as a key reason for the scepticism towards adopting more transparent threshold-based decision rules.[226]

We used data from the Socio-Economic Panel, or SOEP, from 2002 to 2018 (version 35). Fixed-effects and instrumental variable regressions were used to address endogeneity concerns regarding the impact of income on life satisfaction. Our baseline estimates indicate population average monetary valuations of a QALY of €22,717 and €58,533, with and without instrumenting for income. However, alternative specifications and robustness checks lead to varying estimates, highlighting the empirical challenges and the consequences of methodological choices on the obtained monetary values, and areas for future research.

The search for v_Q and the well-being valuation method

Various methods have been used in the ongoing endeavour of obtaining estimates of v_Q, producing a range of conceptually different values. One approach, employed by Mason, Jones-Lee, and Donaldson (2009),[42] bases v_Q on estimates of the value of preventing a statistical fatality, a concept commonly used in public sector safety policies. Another approach calculating v_Q entails using relative risk aversion in relation to income.[43] However, v_Q estimates have predominantly been obtained based on stated preferences, by asking individuals directly about their willingness to pay (WTP) for specific health gains. Ryen and Svensson (2015) summarised the extensive literature that used WTP methods to identify v_Q and reported trimmed mean and median estimates of €74,159 and €24,226 (in 2010 price levels).[38]

Huang et al. (2018) proposed an alternative method for estimating v_Q, based on revealed, although subjective, information: the well-being valuation approach.[44] This method has been applied to obtain monetary valuations for various other non-market goods, including specific health outcomes and diseases,[200,227–229] informal care provision,[230,231] air pollution and natural disasters,[232,233] national security,[234] or the welfare effects of sports events.[235] In their study, Huang et al. (2018) used data from the HILDA panel survey from Australia and obtained v_Q estimates of A\$42,000 (€28,000) to A\$67,000 (€45,000),[44] which were similar to threshold values applied for funding decisions in Australia. Recently, Himmler, van Exel, and Brouwer (2020) applied the wellbeing valuation approach in a cross-sectional sample from the UK to estimate v_Q, as well as an equivalent value for broader well-being. They report a base case v_Q estimate of £30,786 (approximately €35,000).[236]

Both stated preference WTP and well-being valuation approaches have advantages and disadvantages and may answer different questions based on how v_Q is specified. The former allows researchers to tailor their experimental design to specific contexts and control for undesired influences. For instance, WTP can be expressed from an individual or societal perspective,[237] capturing more than self-interested motivations when establishing WTP-based v_Q estimates. Similarly, equity concerns relating to specific health states or streams,[238,239] but also socio-economic health inequalities can be connected with the QALY framework.[240] Furthermore, one can also pose WTP questions from an *ex-ante* or *ex-post* perspective, with the former having the advantage of capturing options value.[241,242] However, the practice of asking individuals directly for the value of a prospect brings unique challenges; hypothetical response bias and insensitivity to scope or framing effects are only some of the practical concerns (see e.g. Kling, Phaneuf, and Zhao(2012))[39] that have been found to apply when obtaining WTP estimates for a QALY.[40,41,154,243]

The well-being valuation approach avoids these challenges by relying on (usually) large-scale observational data, promising to provide a more inclusive picture of the range of preferences over health and wealth across diverse sub-populations. However, the approach limits the scope to respondents' individual *ex-post* valuations, while endogeneity concerns are a prevailing issue as it relies on the estimation of causal effects of health and income to calculate trade-offs.

Methods

Conceptual framework

We generally followed the framework proposed by Huang et al. (2018) for obtaining v_Q.[44] In a simplified model, the subjective well-being (SWB) of individual *i* at time *t*, as a proxy for individual utility, is assumed to be described by:

$$W_{it}=W(Y_{it},\, H_{it}) \tag{18}$$

where W_{it} is a vector of the individual's well-being at all observed time points (w_{it}), Y_{it} is the corresponding incomes (y_{it}), and H_{it} a vector of health states (h_{it}). The total well-being experienced by individual *i* over a time interval of length *T* can then be described by a simple cumulative sum of individual well-being states across time;

$$W_i= \sum_{t=0}^{T} W(Y_{it},\, H_{it}) \tag{19}$$

Within this framework, consider an individual experiencing a change to their health vector ΔH_i within the time window *T*. For the individual to remain on the same level of subjective well-being W_i requires an offsetting income change ΔY_i;

$$W_i= W(Y_i+\Delta Y_i, H_{it}+\Delta H_i) \tag{20}$$

The proposed approach estimates the population average ΔY necessary to offset an imposed hypothetical health state change ΔH over *T* equivalent to one QALY. Therefore ΔY is the compensating income variation for one QALY, or short CIV_{QALY}.

Baseline specification

Following Huang et al. (2018), an ordinary least squares (OLS) fixed-effects regression was estimated to calculate the impact of health and income on SWB within a time window *T* of two years (t_0 and t_{-1}). Modelling SWB as linear despite the cardinal nature of life satisfaction is a widely used approach, see e.g. Ferrer-i Carbonell and van Praag (2002).[227] The underlying empirical model takes the following form;

$$W_{irt}=\alpha+\beta_0 H_{irt}+\beta_1 H_{irt-1}+\delta_0 Y_{irt}+ \delta_1 Y_{irt-1}+\tau X_{irt}+\lambda_i+\mu_r+\epsilon_t+u_{irt} \tag{21}$$

where W_{irt} refers to the subjective well-being of individual *i* living in region *r* at time *t*, measured using life satisfaction data. The individual's health status H_{irt} is captured by health utility values based on the short form six dimensions (SF-6D) instrument and its UK utility tariff.[5] Household income is denoted by Y_{irt}. Lagged variables of health and income were included to not be limited to short-term one-year changes and to partly account for reverse causality. We control for a vector X_{irt} of other potential time-varying confounders. To account for time-invariant unobservables, we incorporated individual

(λ_r), state (μ_r), and time (ε_t) fixed-effects. u_{irt} denotes the error term. Heteroscedasticity-robust standard errors were used in all estimations.

In a second step, we obtained CIV_{QALY} values by dividing the health status coefficients (β_0 and β_1) by the income coefficients (δ_0 and δ_1):

$$\mathrm{CIV}_{\mathrm{QALY}}= \frac{\beta_0+\beta_1}{\delta_0+\delta_1} \tag{22}$$

The corresponding values represent the marginal rate of substitution between income and health with respect to well-being, based on the overall population average. CIV_{QALY} thereby is the empirical conceptualisation of v_Q using the well-being valuation approach. Income outliers (as will be defined below) were dropped from the baseline analysis.

Instrumental variable specification

A well-documented problem of the well-being valuation approach is the endogeneity of the income coefficient estimate. This was frequently addressed using an instrumental variable (IV).[200,228,229] Huang et al. (2018) instrumented income with the occurrence of financial-worsening-events such as personal bankruptcy or large financial losses.[44]

Lacking such information, we followed Luechinger (2009), who used predicted labour-market earnings based on industry-occupation cells as income instrument.[232] The rationale is that shifts in predicted income correspond to industry and/or occupation wide trends, which correlate with the development of negotiated wages or collective wage agreements, but do not reflect individual-level effort or circumstances. Further, it is assumed that the income variance across industries and occupations captures information on the unobserved costs of income generation such as stress and/or associated health risks, and that unobserved selection effects of certain types of individuals into industries and occupations are captured in the time-invariant fixed effects. One advantage of this instrument is that the captured income shifts have a rather permanent nature, whereas financial-worsening-events or lottery wins can be highly transitory shocks. In addition, permanent income shifts have been found to be of higher relevance for individuals' well-being.[244,245]

The identifying assumption is, therefore, that income variation across industries and occupations over time is uncorrelated with individual-level characteristics and especially life satisfaction, besides the effect of income changes themselves. To implement the IV approach, we followed a two-stage least squares estimation procedure. In a first step we estimated the individual's labour market earnings L_{irt} based on the following regression;

$$\mathrm{L}_{\mathrm{irt}}= \alpha+p_0\mathrm{I}_{\mathrm{irt}}+p_1\mathrm{O}_{\mathrm{irt}}+p_2\mathrm{T}_{\mathrm{irt}}+p_3\mathrm{R}_{\mathrm{irt}}+\mu_r+\epsilon_t+u_{\mathrm{irt}} \tag{23}$$

from which we obtained fitted values, constituting the predicted labour earning conditional on the individual's industry-occupation cell (I_{irt} and O_{irt}), work tenure (T_{irt}), and work hours (R_{irt}) and a set of industry- and year-fixed-effects. The obtained predicted labour earnings were summed on the household level and weighted by household composition to obtain the predicted household labour income $\hat{L}_{irt}^{HH}$, the instrument used in the first stage regression;

$$Y_{irt}= \alpha+\bar{\beta}_0 H_{irt}+\bar{\beta}_1 H_{irt-1}+\bar{\delta}_0 \hat{L}_{irt}^{HH}+\bar{\delta}_1 \hat{L}_{irt-1}^{HH}+\bar{\tau} X_{irt}+\bar{\lambda}_i+\bar{\mu}_r+\bar{\epsilon}_t+\bar{u}_{irt} \tag{24}$$

from which we obtained the fitted values for individual income, $\hat{Y}_{irt}$. In the second stage we substituted income Y_{irt} by $\hat{Y}_{irt}$, estimating

$$W_{irt}=\alpha^I+\beta_0^I H_{irt}+\beta_1^I H_{irt-1}+\delta_0^I \hat{Y}_{irt}+\delta_1^I \hat{Y}_{irt-1}+\tau^I X_{irt}+\lambda_i^I+\mu_r^I+\epsilon_t^I+u_{irt}^I \tag{25}$$

The resulting coefficients for health (β_0^I and β_1^I) and income (δ_0^I and δ_1^I) were then included in equation (22) to calculate the IV CIV_{QALY} estimate. For further details please see Appendix A3.

Alternative model specifications

Treatment of outliers

Due to a right-skewed and long-tailed income distribution, with self-reported income often misreported or even exaggerated,[246] income outliers may have a large effect on CIV_{QALY} estimates when using linear models.[247] To identify outliers, which remains challenging for fixed-effects models,[248] we reformulated our base case model as a pooled OLS model and calculated DFbeta, a measure quantifying the impact that dropping an observation has on the coefficient estimate. All observations with a DFbeta larger than 1, the recommended threshold,[249] were dropped from the baseline analysis. In a robustness check we repeated the calculations including these outliers.

Income specification

To accommodate the diminishing marginal return of income we log-transformed income.[209] CIV_{QALY} was then estimated based on a slightly modified equation as used by Olafsdottir, Asgeirsdottir, and Norton (2020) and van den Berg and Ferrer-i Carbonell (2007).[231,250] This entailed dropping the lagged income and health coefficients as used in our base model (equation 22).

$$CIV_{QALY}= \bar{y}^* \left(\exp\left(\frac{-\beta_0 * \frac{1}{\Delta}}{\delta_0} \right) -1 \right) * \Delta \tag{26}$$

In the log-income specification CIV_{QALY} was calculated as the percentage share of annual income (median annual income $\bar{y}$). By construction, CIV_{QALY} values would be confined to be no greater than this income level which may be acceptable when valuing small gains or changes but not a full QALY. Therefore, we added the parameter Δ to the equation and set it to 10. Instead of calculating the monetary equivalent of a one

QALY change we calculated the equivalent of a 0.1 QALY change and multiplied it by 10.

To account for the non-linearity of income without imposing a logarithmic functional form, which may not adequately capture the relationship especially on the lower end of the income distribution, we furthermore tested a piecewise linear specification similar to Olafsdottir et al. (2020).[250] To obtain the appropriate number of income splines and cut-off values, we iteratively combined income-deciles. The equality of coefficient estimates of adjacent splines was tested and non-significantly different splines were gradually combined until coefficients were significantly different and model fit did not improve. CIV_{QALY} values were then calculated for each income spline and also aggregated by weighting according to the number of individuals in the respective splines. Estimating a piecewise IV specification was not feasible, as one distinct income instrument would have been required for each of the splines.

Choice of utility tariff

Lacking a German specific SF-6D utility tariff we relied on the UK time-trade-off based value set to construct health utilities.[5] In an alternative specification we explored the importance of tariff choice by instead applying a recently developed value set from the Netherlands which was estimated using a discrete choice experiment.[122]

Health state dependence of the utility of consumption

Another empirical issue of concern relates to the interaction between health and income and experienced (consumption) utility. This so-called health state dependence implies that the marginal utility gain from a given income change is directly dependent on the underlying health status.[251] So far, there is only inconclusive evidence on the magnitude and the direction of this effect: Finkelstein et al. (2013) found a negative health state dependence, a higher marginal utility of income in good compared to bad health, based on US data. However, replicating their approach using European data, Kools and Knoef (2019) found evidence for positive health state dependence, potentially due to differing provision of public goods in European healthcare systems.[252]

As illustrated by both Finkelstein et al. (2013) and Kools and Knoef (2019), health state dependence has important implications for (health) economic issues such as the optimal design of insurance contracts or individual-level decisions on life-cycle savings. In the context of estimating CIV_{QALY}, which requires a simultaneous measurement of the well-being impacts of both health and income separately, a thorough investigation of the life-cycle development of health states and the associated changes in consumption utility seems warranted.

To explore the potential impact of health state dependence on CIV_{QALY} estimates, we reduced our sample to those individuals that transitioned between health states. Finkelstein et al. (2013) used the onset of chronic diseases for this purpose.[251] While this represents a convenient definition for an elderly population, we took a different approach, allowing us to observe the transition of individuals from good to bad health also for healthier groups. First, we reduced the sample to individuals whose mental or physical short form health questionnaire (SF-12) component scores changed by at least 10, or one standard deviation, throughout their respective observation period (the SF-12 is also used to calculate SF-6D health utilities. Component scores range from 0 (worst) to 100 (best) with a normalised mean of 50 and standard deviation of 10).[253] This was done to ensure that individuals in this group have experienced a consequential change in their mental and/or physical health. Good health states were defined as periods in which either of the two scores was above their respective individual-level mean; bad health states if they were below. Secondly, we conditioned on the consecutive observation of differing health states with at least two consecutive periods needed to be observed in either state. This allowed us to estimate CIV_{QALY} for good and bad health separately while also ensuring that individuals transition into longer-term health states (see Appendix A4 for details). Importantly, the sample included individuals transitioning from good to bad health and vice versa, although the former is most frequent.

Data

We used data from the annual SOEP panel survey, providing a representative sample of the adult (aged 16+) German population.[254] Ethical approval with respect to the surveying process generating the underlying data was obtained by the SOEP researchers directly. SF-6D health utilities were constructed from SF-12 data, which is biennially included in the survey since 2002. To facilitate the specified two-year time-frame T used for the CIV_{QALY} calculations, and to prevent dropping observations from every second year, we linearly imputed SF-6D values for intermediate years. However, this was only done if individuals were observed for three consecutive years with two completed SF-12 surveys.

Life satisfaction was measured on a 10-point scale ranging from 0 (*"completely dissatisfied"*) to 10 (*"completely satisfied"*). Information on individuals' income was based on self-reported monthly net household income. To account for differences in household composition, we calculated equivalised household income, following the definition by.[255] Income data was converted to 2018 prices using the official consumer price indices.[256]

To construct our instrument, we extracted information on net labour income and individuals' industry and occupation. We dropped households with individuals where information on labour income but not on industry/occupation was available. Predicted labour income was assumed to be zero for all individuals with no labour income information, or who stated that they were not employed. Following Luechinger (2009) we added a constant of €1 to all incomes for the log-income specification.[232]

We furthermore extracted information on a similar set of variables as used by Huang et al. (2018) to control for confounding factors.[44] These included age, disability, marital status, employment status, educational attainment and leisure time. Table 19 summary statistics of the analysis data, consisting of 29,735 individuals providing 186,906 individual-year observations. Appendix Table A1.1 provides an overview of the conditioning applied to the SOEP data, while Appendix Table A1.2 shows that the sub-sample of employed individuals who were dropped because of missing industry/occupation information is comparable to the remaining sample of employed individuals. As the exclusion of individuals without at least two consecutive SF-6D values was the only major selection criterion, the sample remained largely representative for the overall German population.

Table 19: Descriptive statistics

Variable	Mean	Std. Dev.	Description
Life satisfaction	7.09	1.71	0 (lowest) to 10 (highest)
Income in 1000€	2.03	1.29	Monthly household income in e
SF-6D utility	0.73	0.13	0.345-1, 1 perfect health
Disability	0.14	0.35	1 if disability status
Age in years	53.67	15.78	
(de facto) Married	0.67	0.47	1 if married, living together
Education: Primary	0.12	0.32	1 if primary educated
Education: Secondary	0.63	0.48	1 if secondary educated
Education: Tertiary	0.25	0.43	1 if tertiary educated
Leisure time	2.18	2.03	Hours per day
Employed	0.56	0.50	1 if employed
Unemployed	0.04	0.21	1 if unemployed
Work hours	21.22	20.99	Hours per week
Tenure	7.03	9.96	Years at current job
Individuals * Years		186,902	
Individuals		29,735	

Note: Own calculations based on SOEP Waves 2002-2018.

Results

Baseline results

The baseline OLS and IV results, are shown in Table 20, separating between results using the full dataset with imputed SF-6D values, and the dataset without imputation. To construct our instrumental variables, we predicted labour incomes based on industry/occupation for 125,229 observations. Appendix A3 provides details on this prediction and the associated errors, which were small for the largest part of the income distribution. The instruments were significant in the first stage regression (Appendix Table A3.1) and passed the Cragg-Donald weak identification test (F-value: 1,864 and 192). This indicates a high relevance of the instrument, a common finding for this type of instrument.[232,244] The Hausman test for endogeneity of the instrumented variables was significant, signalling that income should not be treated as exogenous.

Equivalised monthly household income, health status (SF-6D utility), and their lagged values were positive and significant predictors of life satisfaction in the OLS specification. This was also the case when instrumenting for income, except that the lagged income coefficient was insignificant. We observed a two-fold increase in the income coefficients in the IV model (0.048 vs. 0.098), a similar magnitude to what has been observed in previous studies using the SOEP.[244,257] Interestingly, the difference is minimal compared to what was observed by Huang et al. (2018),[44] who reported an IV coefficient which was 130 times larger than the OLS coefficient (0.080 and 0.0006). Applying the estimated income and SF6D coefficients to equation (22) resulted in a CIV_{QALY} value of €58,533 in the OLS model and €22,717 when instrumenting for income. This value represents the average amount of additional income necessary to maintain the same level of life satisfaction if a hypothetical health change of one QALY is imposed.

Without SF-6D imputation, reducing our sample to 85,433 observations across 21,718 individuals, the OLS results increased by a factor of 1.38 to €80,522 while the IV-based value increased by a factor of 1.24 to €28,130. These differences were driven by larger SF-6D and income coefficients compared to the baseline calculations, possibly resulting from increased within-person variance and time-frame T being two years instead of one. For the remainder of the results presented, we will be using the full dataset with imputed SF-6D values to make use of the largest amount of information available.

Table 20: Baseline results

	SF-6D imputation				No imputation			
	OLS		IV		OLS		IV	
Income in 1000€	0.05***	(0.01)	0.10***	(0.03)	0.05***	(0.01)	0.14***	(0.05)
Income (*t* – 1)	0.01	(0.01)	0.04	(0.03)	-0.00	(0.01)	-0.00	(0.07)
SF-6D utility	3.12***	(0.06)	3.12***	(0.05)	3.52***	(0.06)	3.51***	(0.05)
SF-6D utility (*t* – 1)	0.10*	(0.06)	0.10*	(0.05)	0.47***	(0.05)	0.46***	(0.05)
disability	-0.14***	(0.02)	-0.14***	(0.02)	-0.09***	(0.03)	-0.09***	(0.02)
Age	0.09***	(0.01)	0.08***	(0.02)	0.05***	(0.01)	0.05***	(0.01)
Age squared	-0.00***	(0.00)	-0.00**	(0.00)	-0.00**	(0.00)	-0.00	(0.00)
(de facto) Married	0.18***	(0.02)	0.18***	(0.02)	0.17***	(0.03)	0.16***	(0.02)
Primary education	-0.18*	(0.09)	-0.21***	(0.08)	-0.10	(0.15)	-0.13	(0.13)
Tertiary education	-0.18***	(0.06)	-0.19***	(0.05)	-0.19***	(0.07)	-0.20***	(0.07)
Leisure time	0.03***	(0.01)	0.03***	(0.00)	0.03***	(0.01)	0.03***	(0.01)
Leisure time squared	-0.00***	(0.00)	-0.00***	(0.00)	-0.00**	(0.00)	-0.00***	(0.00)
Unemployed	-0.52***	(0.03)	-0.53***	(0.02)	-0.53***	(0.04)	-0.53***	(0.03)
Work hours	0.00***	(0.00)	0.00***	(0.00)	0.00***	(0.00)	0.00	(0.00)
Tenure	-0.01***	(0.00)	-0.01***	(0.00)	-0.01***	(0.00)	-0.01***	(0.00)
Model statistics								
Cragg-Donald			1,864				192	
Anderson			3,642				382	
Endogeneity test			10.0				5.8	
BIC	540,754		540,995		250,099		236,538	
Observations	186,902		186,902		93,450		85,433	
Individuals	29,735		29,735		29,735		21,718	
CIV/QALY in €	58,533		22,717		80,522		28,130	

Note: Own calculations based on SOEP Waves 2002-2018; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. BIC Bayesian information criteria.

Table 21 columns 2-3 contains estimates for East and West Germany separately, motivated by the persisting differences in life satisfaction and income levels.[258,259] OLS-based CIV_{QALY} estimates were €75,748 in the West and €28,548 in the East. The IV-based estimate was also higher in the West compared to the East (€20,750 and €12,982), although the relative difference was lower (factor of 3.64 and 2.20). In both models, this difference was mainly driven by a considerably larger income coefficients in the East, likely due to the prevailing income differences between West and East; observed average monthly equivalised income was €2,140 in the West and only €1,652 in the East.

We investigated the (undesired) impact of macro-economic conditions on CIV_{QALY} estimates by excluding the years of the financial crisis and recession in Germany (2007-2009). As shown in Table 22 (columns 4-6), this had only a minor impact on the OLS and IV CIV_{QALY} values (€54,567 and €20,574). However, estimates based on the pre-crisis time periods 2002-2006 (€56,640 and €7,720) were substantially lower compared to estimates based on data from 20102018 (€70,572 and €24,811). This resulted from larger estimated effects of income in earlier periods, which may both be a result of a positive trend in incomes or a shift in population preferences and values over the last decades. Appendix Table A2.1 provides further results on age and gender subgroups.

Table 21: Results by region and time period

	East		West		w/o 2007-2009		2002-2006		2010-2018	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Income in 1000€	0.13***	0.18**	0.04***	0.07**	0.05***	0.11***	0.06***	0.29***	0.04***	0.09*
	(0.02)	(0.08)	(0.01)	(0.04)	(0.01)	(0.04)	(0.01)	(0.09)	(0.01)	(0.05)
Income (t – 1)	0.00	0.03	0.01	0.04	0.01*	0.05	-0.00	0.10	0.01	0.04
	(0.02)	(0.06)	(0.01)	(0.03)	(0.01)	(0.03)	(0.01)	(0.08)	(0.01)	(0.04)
SF-6D utility	2.90***	2.90***	3.18***	3.17***	3.16***	3.15***	2.93***	2.92***	3.08***	3.08***
	(0.13)	(0.12)	(0.07)	(0.07)	(0.07)	(0.07)	(0.15)	(0.15)	(0.08)	(0.08)
SF-6D (t – 1)	-0.12	-0.12	0.16**	0.16**	0.10	0.09	0.06	0.06	-0.07	-0.07
	(0.12)	(0.12)	(0.07)	(0.07)	(0.07)	(0.07)	(0.14)	(0.14)	(0.08)	(0.08)
Model statistics										
Cragg-Donald		323.9		680.2		783.4		181.2		494.3
Anderson		544		1,266		1,430		329		907
Endogeneity test		1.5		5.8		9.7		8.2		2.7
BIC	127,072	127,092	412,723	412,877	431,238	431,487	129,869	130,432	276,374	276,464
N	43,447	43,447	143,361	143,361	151,461	151,461	48,678	48,678	101,048	101,048
CIV/QALY in €	20,750	12,982	75,748	28,548	54,567	20,574	56,640	7,720	70,572	24,811

Note: Own calculations based on SOEP Waves 2002-2018; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. BIC Bayesian information criteria.

Specifications related to income

Re-estimating our baseline models including four individual-year observations which were flagged as outliers lead to a considerably lower income coefficient in the OLS model (Table 22 columns 3-4). This increased the CIV_{QALY} value to €82,484. The IV estimates were only minimally affected by this (€22,782). The outlier observations corresponded to two individuals from the same household, which reported a drop in monthly income from €142,534 to €14,051 within two consecutive years, while reporting constant life satisfaction.

In the models using log-transformed income (Table 22 columns 5-6), the income coefficient was 0.24, larger than reported before by Pischke (2011) (0.125 to 0.182), who also used the SOEP.[257] The corresponding IV coefficient, with a value of 0.63, was on the higher end of previous IV estimates based on the industry-wage structure and the SOEP: Luechinger (2009) reported an estimate of 0.55,[232] while Pischke (2011) reported values ranging from 0.489 to 0.617.[257] Previous estimates based on instruments using lagged or future income shocks were also similar, with Bayer and Juessen (2015) providing a range of 0.45 to 0.50 for permanent income shifts. Bayer and Juessen (2015) used only data from West Germany, possibly leading to a downward bias due to higher income levels in the West. Similarly, both Pischke (2011) and Luechinger (2009) use SOEP waves from the years before the East German SOEP sample was established in 1990 alongside waves containing samples from both former German states past 1990.[244] The log-transformation resulted in considerably larger CIV_{QALY} values compared to the baseline. The OLS values increased by a factor of 2.63 to €153,877 while the IV values increase by a factor of 3.59 to €81,649. Huang et al. (2018) did not observe a large difference between linear and log income based estimates. However, they multiplied the ratio of income and health coefficients as in equation (22) with the median income to obtain CIV_{QALY} (as opposed to equation (26)).[44]

Table 22: Income specifications

	Baseline		Without outliers		Log income		Piece-wise
	OLS	IV	OLS	IV	OLS	IV	OLS
Income in 1000€	0.05^{***} (0.01)	0.10^{***} (0.03)	0.03^{***} (0.01)	0.10^{***} (0.03)			
Income in 1000€ (t – 1)	0.01 (0.01)	0.04 (0.03)	0.01^{***} (0.00)	0.04 (0.03)			
SF-6D utility	3.12^{***} (0.06)	3.12^{***} (0.05)	3.12^{***} (0.06)	3.12^{***} (0.06)	3.18^{***} (0.05)	3.16^{***} (0.05)	3.18^{***} (0.05)
SF-6D utility (t – 1)	0.10^{*} (0.06)	0.10^{*} (0.05)	0.10^{*} (0.06)	0.10^{*} (0.06)			
Log income					0.24^{***} (0.02)	0.63^{***} (0.13)	
1st income spline							0.43^{***} (0.05)
2nd income spline							0.27^{***} (0.05)
3rd income spline							0.11^{***} (0.02)
4th income spline							0.01 (0.01)
Model statistics							
Cragg-Donald		1,863.7		825.8		1,329.9	
Anderson		3,642.0		1,529.4		1,278.2	
Endogeneity test		10.0		12.9		9.7	
BIC	540,755	540,995	540,801	541,306	540,506	541,501	540,448
Observations	186,902	186,902	186,906	186,906	186,902	186,902	186,902
CIV/QALY €	58,533	22,717	82,484	22,782	153,877	81,649	97,486
w/o 4th spline	540,755	540,995	127,072	127,092	412,723	412,877	19,515

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. BIC Bayesian information criteria. Instrumental variable did not pass weak identification tests for piecewise income specification. CIVs for piecewise regression represents population-weighted averages of all splines or the first three splines (€7,347, €11,686, €29,548, and €409,810).

The piecewise linear specification was estimated with ultimately four income splines. The cut-off points were at the 20th percentile (€1,200), the 40th percentile (€1,546), and the 80th percentile (€2,635). Figure 18 plots the overall distribution of life satisfaction across income, and the linear fit of life satisfaction across splines, indicating a non-linear, diminishing pattern. The spline specific CIV_{QALY} values were €7,347, €11,686, €29,548, and €409,810. The population aggregated CIV_{QALY} was €97,486. This estimate was driven by the large CIV_{QALY} value in the fourth income spline, where the income coefficient was insignificant. Using the three significant splines lead to a CIV_{QALY} value of €19,515.

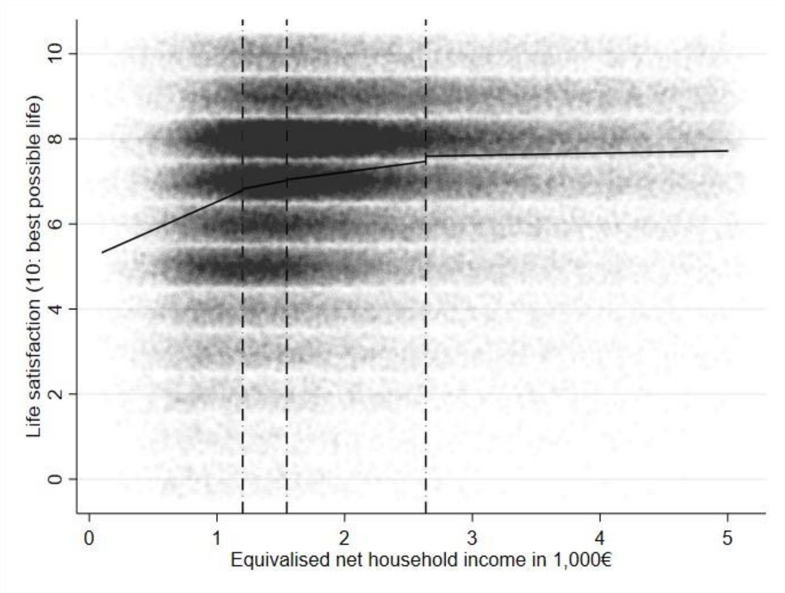


Figure 18: Relationship between life satisfaction and income across income splines. Note: Life satisfaction values are depicted as small grey dots. Black dash-dotted vertical lines represent the income splines used in the piece-wise linear regression. Black horizontal lines plot the linear fit within these splines.

Specifications and issues related to health

Choice of SF-6D value set

Applying the Dutch SF-6D value set shifted the distribution of health utilities (Figure 19), with the mean utility decreasing from 0.725 to 0.554. These differences likely reflect methodological differences rather than actual variation in health state preferences between the UK and the Netherlands as UK and Dutch tariffs for the EQ-5D have been shown to be similar.[260]

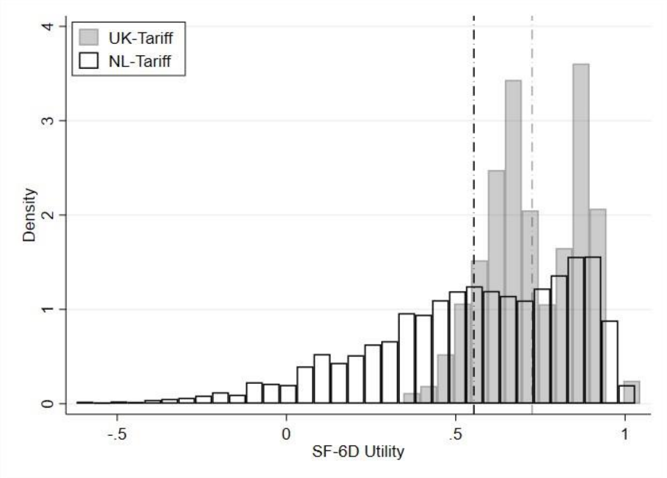


Figure 19: SF12 index values using UK and Dutch tariffs. The black dash-dotted line indicates the Dutch tariff mean. The grey dash-dotted line indicates the UK tariff mean. The distributions and means reflect SF-6D values based on self-reported SF12 questionnaires only.

The estimated CIV_{QALY} values using the Dutch SF-6D tariff were markedly smaller (Table 23). The OLS estimates decreased from €58,533 to €32,534, while the IV estimates decreased from €22,717 to €13,054. This shift was caused by the smaller SF-6D coefficients (3.12 to 1.78), resulting from the wider spread of the Dutch tariff, which ranges from -0.44 to 1, allowing for negative health state utility, instead of 0.345 to 1 as in the UK value set. The same actual change in health corresponds to a larger change in SF-6D utility in the Dutch tariff which reduces the impact of a (hypothetical) one unit change in SF-6D on life satisfaction.

Table 23: Choice of SF-6D tariff

	UK tariff		Dutch tariff	
	OLS	IV	OLS	IV
Income in 1000€	0.05***	0.10***	0.05***	0.09***
	(0.01)	(0.03)	(0.01)	(0.03)
Income in 1000€ (t – 1)	0.01	0.04	0.01	0.05*
	(0.01)	(0.03)	(0.01)	(0.03)
SF-6D utility	3.12***	3.12***	1.78***	1.78***
	(0.06)	(0.05)	(0.03)	(0.03)
SF-6D utility (t – 1)	0.10*	0.10*	0.05	0.05
	(0.06)	(0.05)	(0.03)	(0.03)
Model statistics				
Cragg-Donald		1,863.7		825.8
Anderson		3,642.0		1,529.4
Endogeneity test		10.0		12.9
BIC	540,755	540,995	540,801	541,306
Observations	186,902	186,902	186,906	186,906
CIV/QALY €	58,533	22,717	82,484	22,782
w/o 4th spline	540,755	540,995	127,072	127,092

Note: * *p < 0.10, ** p < 0.05, *** p < 0.01*. BIC Bayesian information criteria.

Health state dependence of the utility of consumption

We explored the potential impact of health state dependence on CIV_{QALY} estimates by restricting our sample to individuals experiencing a substantial health change and splitting their respective observation periods into good and bad health states. The resulting sample was considerably smaller, including only 5,112 individuals yielding 48,861 observations. Nevertheless, the summary statistics suggests that the sample is still comparable to the full population sample (see Appendix Table A4.1). Table 24 depicts the corresponding estimation results. Compared to the baseline estimates using the full sample, CIV_{QALY} values based on the combined good and bad health state samples were lower in the OLS model (€39,482) and similar in the IV specification (€20,377). For "good health states", the corresponding CIV_{QALY} estimates were lower with €33,336 and €16,532. For "bad health states", the OLS-based CIV_{QALY} estimate was €38,374 and the IV-based estimate €11,779.

Table 24: Health state dependence

	UK tariff		Dutch tariff		Bad health	
	OLS	IV	OLS	IV	OLS	IV
Income in	0.07***	0.17**	0.05***	0.11	0.08**	0.32
1000€	(0.01)	(0.07)	(0.02)	(0.08)	(0.04)	(0.24)
Income	0.03**	0.02	0.03**	0.05	0.03	0.05
(*t* – 1)	(0.01)	(0.06)	(0.01)	(0.06)	(0.03)	(0.17)
SF-6D utility	3.62***	3.60***	2.51***	2.50***	4.10***	4.03***
	(0.11)	(0.09)	(0.14)	(0.12)	(0.38)	(0.37)
SF-6D utility	0.10	0.11	0.12	0.12	0.32	0.32
(*t* – 1)	(0.10)	(0.10)	(0.12)	(0.11)	(0.26)	(0.27)
Model statistics						
Cragg-Donald		620.7		425.1		95.9
Anderson		1,208.4		828.1		188.4
Endogeneity test		3.0		1.8		1.0
BIC	150,481	150,558	102,463	102,497	37,832	37,899
Observations	48,861	48,861	35,401	35,401	13,460	13,460
CIV/QALY €	39,482	20,377	33,336	16,532	38,374	11,779

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. BIC Bayesian information criteria.

Important to note is that the drop in the IV based results for the bad health state primarily resulted from a larger income coefficient estimate, even though the SF-6D coefficients increased considerably. These results indicate that there is a positive health state dependence of income in line with the results for Germany by Kools and Knoef (2019).[252] Unfortunately, we were not able to follow Kools and Knoef (2019) and Finkelstein et al. (2013) in focusing on nonworking individuals to ensure stable income across health states, ruling out that the increased income coefficients are driven by individuals losing their income, and hence having a larger marginal utility of additional earnings.[251] For our analysis, such a restriction was not feasible, as within-person income variation is necessary to estimate the income coefficients. However, the general empirical pattern remains the same when excluding individuals with large negative income differences between health states (see Appendix Table A4.2). This also holds when only considering the working population (Table A4.3) and those experiencing sudden and severe health changes (Table A4.4).

Robustness checks

Lastly, we tested the robustness of our baseline results to some general concerns regarding our estimation strategy (Table 25). In a first robustness check, we limited our sample to individuals which were in paid employment and provided industry-occupation information, the same sample which was used to obtain estimates for predicted labour income for the IV regression. The resulting OLS-based CIV_{QALY} was slightly lower than the baseline at €52,829, while the IV-based value was slightly higher than the baseline at €26,097. These differences were driven by the smaller SF-6D coefficients in both OLS and IV models, likely resulting from the working population being healthier as individuals without labour income (the unemployed and retired). The sum of both income coefficients was smaller in the corresponding IV-calculations compared to baseline, increasing the CIV_{QALY}.

Next, we followed Luechinger (2009) by excluding households with self-employed main income earners, as the income measurement error was likely to be amplified among these individuals.[232] Self-employed individuals are often reluctant to disclose their income, while also experiencing unstable income streams and hence, even if not reluctant to report, they might simply misreport accidentally. The resulting CIV_{QALY} estimates and income and SF-6D coefficients were similar to the baseline estimates (€55,359 and €20,352).

Another concern relating to the instrument is that observed income changes may also relate to individual effort, which likely impacts income differently across industries and occupations. Unfortunately, effort cannot be observed. To nevertheless explore this, we use information on reported bonuses, gratifications, or profit sharing to identify the group of individuals for whom this might be a relevant concern, as for them effort would have the highest impact on income and life satisfaction. To test the robustness of our results to this potential bias, we estimate our baseline models excluding such observations. The results in Table 25 columns 7-8 suggest that this bias is relatively limited.

To investigate the potential impact of dropping employed individuals without industry/occupation information (as required for constructing the IV), we included those observations in a further robustness check (last column 7). The corresponding OLS estimates for income coefficients and CIV_{QALY} (€62,266) are comparable to our baseline estimates. However, by construction, we cannot confirm this for the IV estimates.

Table 25: Robustness check

	Baseline		Working only		No self-employed		No bonus income		Ind/occ
	OLS	IV	OLS	IV	OLS	IV	OLS	IV	OLS
Income in	0.05***	0.10***	0.05***	0.05	0.07***	0.05	0.05***	0.14***	0.04***
1000€	(0.01)	(0.03)	(0.01)	(0.03)	(0.01)	(0.04)	(0.01)	(0.04)	(0.01)
Income	0.01	0.04	0.01	0.07**	0.00	0.08**	0.01	0.02	0.01**
(t – 1)	(0.01)	(0.03)	(0.01)	(0.03)	(0.01)	(0.03)	(0.01)	(0.03)	(0.01)
SF-6D	3.12***	3.12***	2.95***	2.94***	2.97***	2.97***	3.12***	3.11***	3.14***
utility									
	(0.06)	(0.05)	(0.08)	(0.07)	(0.08)	(0.07)	(0.06)	(0.06)	(0.06)
SF-6D	0.10*	0.10*	0.07	0.06	0.01	0.01	0.10*	0.11*	0.12**
(t – 1)	(0.06)	(0.05)	(0.07)	(0.07)	(0.08)	(0.07)	(0.06)	(0.06)	(0.06)
Model statistics									
Cragg-Donald		1,864		1,356		1,898		719	
Anderson		3,642		2,638		3,633		1,334	
End. test		10.0		5.4		7.4		10.1	
BIC	540,755	540,995	319,169	319,323	279,896	280,043	502,827	503,172	578,002
N	186,902	186,902	116,125	116,125	101,703	101,703	172,998	172,998	198,950
CIV/QALY in €	58,533	22,717	52,829	26,097	44,058	21,382	53,974	20,464	62,266

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. BIC Bayesian information criteria. Ind/occ refers to specification where individuals without industry/occupation information were included.

Discussion

Applying the well-being valuation approach to longitudinal health and income data from Germany, we estimated the monetary equivalent value of one year in full health v_Q (equivalent to one QALY). Beyond demonstrating the feasibility of this approach in a new country context, we explored additional empirical and methodological challenges with implications for the practical usefulness of well-being valuation based v_Q estimates (denoted as CIV_{QALY}).

Overview and context of results

Figure 20 presents an overview of our CIV_{QALY} estimates. The baseline calculations provided average monetary valuations of a QALY of €58,533 (OLS) and €22,717 (IV). CIV_{QALY} estimates varied across model specifications with the bulk of values lying between €20,000 and €60,000 and the (OLS) log-income specifications reaching the maximum value of €153,877. Instrumenting for income consistently lead to lower values, a common finding in the well-being valuation literature.[250]

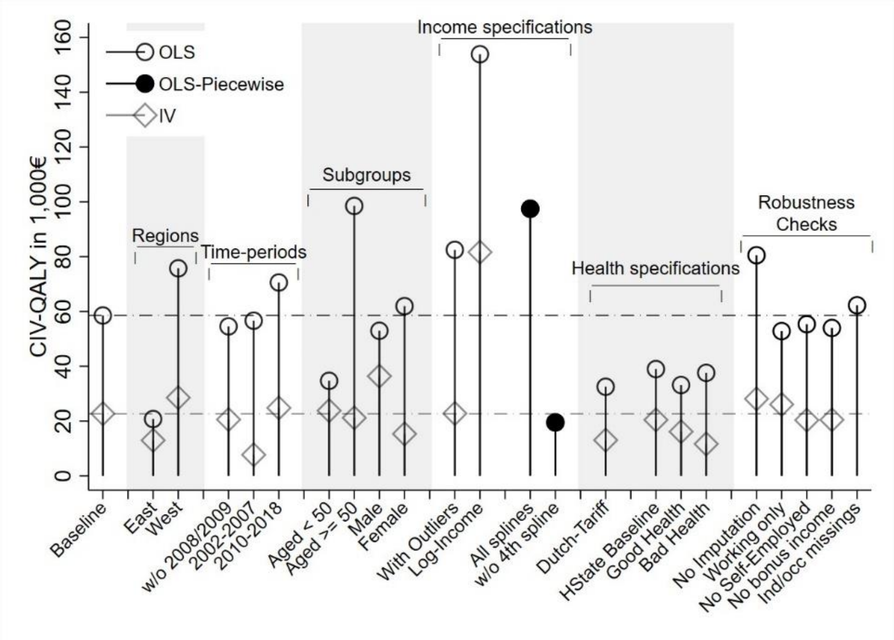


Figure 20: Overview of CIV_{QALY} estimates. Note: The horizontal dash-dotted lines indicate our baseline CIVQALY estimates from the baseline OLS (black) and IV (grey) specifications.

The range of CIV_{QALY} estimates obtained in our study fit into the ballpark of more reasonable stated preference estimates.[38] Furthermore, it is important to note that all IV CIV_{QALY} estimates, except the log-income specification, fell within the range of v_Q estimates for Germany of €4,988 to €43,115 reported by Ahlert et al. (2016), who provided the only v_Q estimates until now.[40] A first approximation of an *opportunity cost* *based QALY threshold value*, or k_Q, for Germany was reported by Woods, Revill, Sculpher, and Claxton (2016).[261] Using empirical estimates of health care opportunity costs for Germany, and the relationship between GDP per capita and the value of a statistical life, they calculated a k_Q range of €19,276 to €24,374 (in 2018 euros). A recent related study by Ochalek and Lomas (2020) reported estimates of cost per DALY averted (essentially the reciprocal of a QALY gain) for Germany of €47,116 to €74,650 (in 2018 euros).[262]

Limitations and strengths of the analysis

IV-based estimates rely on restrictive assumptions relating to their unbiasedness and informational value. A valid concern is that occupational choice may be related to other unobserved confounders, such as personality traits or income preferences.[263] The use of individual fixed-effects should somewhat alleviate such concerns due to the rather stable nature of personality traits,[264] but they cannot provide complete assurance. A further assumption is that being employed in a certain industry/occupation should not have a significant, direct effect on life satisfaction, therefore violating the exclusion restriction. Appendix Tables A3.6 and A3.7 show that, controlling for income and other confounders, this effect is not zero, but modest and mostly insignificant. One additional drawback that is rarely explicitly discussed but of great importance in the well-being valuation context, is that IV estimates only yield a local average treatment effect.[265] Using predicted labour income as an instrument, at least questions the generalisability of our IV estimates to the full, also non-working, population. Further, as we are not able to address all sources of measurement error with respect to income, the remaining upward bias in the income coefficients would imply a downward bias in the estimated CIV_{QALY} values.

In addition, income variation in industry-occupation cells predominantly consists of *positive, upward shifts in wages* (and differences therein). This is conceptually different to financial worsening events, as used by Huang et al. (2018), as these capture *income losses*.[44] Ambrosio, Clark, and Zhu (2018) report a persistent direct effect of financial worsening (and improvement) events on life satisfaction beyond income-changes, raising concerns on the general appropriateness such events as income instruments.[266] In any case, given income loss aversion,[267] our IV based CIV_{QALY} estimates likely represent a lower-bound.

The potential endogeneity of health (status) in life satisfaction regressions due to reverse causality,[79,268] which is rarely addressed in the related literature, is a further limitation. This endogeneity could be addressed by appropriate instruments or identifying health shocks which are plausibly exogenous, such as heart attacks or strokes. However, besides practical issues like data availability, it is questionable how generalisable such localised causal effects would be for the overall impact of the multi-dimensional construct of health on life satisfaction. Heterogeneity may exist both concerning the type of health shocks, but also relating to their timing within the (life cycle) health distribution. Whether or not our estimates of the impact of health are biased upwards or downwards can therefore not be easily ascertained. In the one previous article in the related literature that addressed endogeneity directly, Brown

(2015) found that the health coefficient was slightly overestimated when not instrumented.[200] Assuming this also holds in our context, this would imply that there is an upward bias in our CIV_{QALY} values resulting from the endogeneity of health.

A more practical limitation relating to measuring health was that we had to impute SF-6D utilities for every second year to make full use of the SOEP's rich annual data. This required us to condition the sample on individuals who had at least three consecutive observations, which may have resulted in underestimating the impact of deteriorating health, since individuals are more likely to discontinue their participation in a longitudinal survey following a negative health shock.

A final limitation lies in the potential presence of double-counting as subjective well-being enters the model twice: As an implicit consideration in the SF-6D health state valuation tasks (on which the scoring of our health measure is based on), and as a proxy for experienced utility (equation 18). To what extent this is problematic is difficult to assess. To avoid this double counting, one could use an unweighted sum score of the SF-6D levels. However, this raises the question of the appropriate anchoring. Using such a sum score, rescaled to a 0 to 1 range (expanding the number of levels of the first two SF-6D dimensions to five to not impose any weighting) lead to lower CIV_{QALY} estimates in the unimputed dataset (Appendix Table A2.2, columns 4-5). However, when imposing the same anchor and therefore range as in the original SF-6D tariff (0.345 to 1), the OLS and IV results (€88,867 and €30,567) were much closer to the unimputed baseline estimates (€80,671 and €27,777).

It seems that not the differential weighting between the dimensions caused the larger differences, but the different anchors, i.e., the lowest utility. Another alternative approach entailed eliciting *CIV* values for different dimensions directly by regressing on all levels of the SF-6D, which did not impose any weighting. Adding up the resulting *CIV* values of the lowest level of all six dimensions, summed up to a cumulative value of moving from the best possible to the worst possible health state of €79,013 and €27,489, which again resembled the unimputed baseline estimate (Table A2.2). While these sensitivity checks somewhat alleviate the concerns about double-counting, the latter revealed that 46% of the CIV_{QALY} value stemmed from the impact of mental health on life satisfaction. It is likely that the mental health dimension also plays a dominant role in our baseline calculations. Whether this in itself is problematic lies outside the scope of this paper, as it relates to a more general issue of the well-being valuation approach: is life satisfaction the best (available) proxy for experienced utility?

Implications of findings

There are several practical implications of our study for future applications of the well-being valuation approach in general, and its use for estimating v_Q in particular. First, judging from the impact outliers have in the OLS specification (Table 22), subsequent applications of the approach using linear models should report on the occurrence and treatment of outliers. Secondly, given that the functional form of income had a large impact on our estimates its final specification has to be well argued and reporting results for other alternative functional forms seems warranted. The piecewise linear specification seems to be a promising alternative, given that it is more flexible and gives all income groups a proportional weight. This approach, however, comes at the price of increasing the number of variables that need to be instrumented for.

Third, the choice of utility tariffs for the health instrument matters greatly. Especially the range of the scoring algorithm has a large impact (Table A2.2), as an imposed one unit change in health utility implies a different change in health if the range goes from 0.345 to 1 or -0.44 to 1. How to overcome this issue while facilitating cross-country comparisons and how this relates to the underlying QALY concept, should further be discussed in future applications. Lacking country specific tariffs, it may be convenient to opt for a tariff whose origin can be placed in cultural and socio-economic proximity to the country to be investigated. However, the impact of methodological peculiarities in how these tariffs were generated are relevant. It would have been interesting also to compute CIV_{QALY} estimates based on the more widely used EQ-5D health utilities and compare the implications of differences in scope and range of the health instrument used on CIV_{QALY} values. Unfortunately, EQ-5D is rarely included longitudinal surveys. Lastly, the differing values obtained when considering East and West Germany separately, or specific time periods (Table 21), also highlight the potential importance of the specific country context for CIV_{QALY} calculations.

One of the major conceptual issues discussed in our analysis, with direct relevance for the practical value of any empirically estimated *CIV* of health, is the health state dependence of utility. We attempted to provide indicative evidence on how health state dependence might affect estimated CIV_{QALY} values. However, it remains unclear whether empirical approaches based on self-reported (panel) data can produce reliable estimates if health state dependence is prevalent and survey participation and attrition is (partially) driven by health changes over time. We found considerable differences in the estimated CIV_{QALY} values when comparing periods of good and bad health within individuals (Table 24). As the underlying point estimates depicted substantial uncertainty, these findings should be interpreted with caution and merely as indicative evidence for the role of health state dependency in this context. The impact of this sub-sample of individuals on the population wide CIV_{QALY} value is likely small, as attrition is high once individuals experience bad health states, long-term or very severe health shocks. Hence, a pragmatist might argue that this issue is of theoretical interest only. We would argue, however, that this is an inherent limitation of self-reported observational data and its *ex-post* perspective in this context. Stated preference methods would allow for an explicit *ex-ante* consideration of this issue through tailored sampling strategies and survey design.

An additional conceptual concern related to health state dependence is the question of adaptation to bad health over time.[44] Adaptation implies the gradual return of subjective well-being to pre-health-shock levels despite continued (or deteriorating) bad health.[269] This phenomenon has been documented before using the SOEP data and would generally decrease estimated CIV_{QALY},[270] as the marginal utility of health would decrease with time spent in bad health. To what extend this represents an estimation error, however, is debatable and depends on what is perceived to be the *"true"* impact of ill-health on well-being over time, and whether adaptation, if present, should be corrected for. The recent findings by Etilé, Frijters, Johnston, and Shields (2020),[271] who documented a heterogeneous distribution of adaptive potential across subgroups, underline the relevance of this concern also from a normative perspective.

The previous remarks highlight avenues for future research, like investigating the causal effect of health on life satisfaction, for example using instrumental variable regressions. In addition, the approach would crucially benefit from further research into the impact of income on life satisfaction, for example using (natural) experiments. The regular inclusion of variables that represent valid instruments for income into different population panel surveys could also be beneficial for further exploring the reliability and validity of these instruments and the approach as a whole, as it would allow cross-national replications of results. Meanwhile, future applications may draw upon recent advances into the generalisability of IV-based estimates to explore how these concerns can be addressed within the framework of available instruments.[272] Further, linking survey data on individual-level subjective well-being measures with detailed administrative records on income, health, and care consumption would also be a fruitful direction for further inquiry, resolving some of the enumerated concerns. With respect to the question of health state dependency, for example, it would be possible to determine the extent to which survey data has an inherent blind spot due to the attrition of individual following severe health shocks. In addition, such data could also be used to explore a wider range of specification choices within the general empirical strategy used, for example with respect to the choice of control variables. Here, we deliberately followed Huang et al. (2018),[44] as the set of basic control variables they propose is available in most national panel surveys, which facilitates replications across country-contexts. However, there is ample room for extending the analysis by considering a wider set of control variables and their impact on CIV_{QALY} estimates, or even to altogether choose a different approach such as shrinkage estimators (e.g., LASSO) or matching to address endogeneity concerns around the impact of health and/or income on life satisfaction.

A final issue concerns the practical application of our v_Q estimates. If certain (health) policies/interventions in Germany were to be evaluated using a v_Q value from our study, which range from around €20,000 (IV) to €60,000 (OLS), we have to highlight the following: Health care funding decisions in Germany are not based on cost utility analysis, partially because thresholds were considered to be difficult to define.[226] Finding comparable monetary estimates using a compensating income variation and stated preference studies to some extent puts this into perspective. However, our study cannot provide a definite answer regarding which estimate is most accurate to be used in different contexts. This relates to the uncertainty surrounding these estimates and the underlying assumptions, but also to normative or distributional questions, which need to be addressed in the future.[273] While our piecewise regression results somewhat reflect such concerns by constructing v_Q estimates using a weighted mean of the different parts of the income distribution, this is only a first, very simplistic approach. When used in a normative context, like decisions on reimbursement of technologies, explicit policy (debate and) support is required. Applied studies could use the range we provided to highlight the impact of varying v_Q estimates on their results and recommendations, keeping in mind that for specific sub-populations our v_Q estimates might not be directly applicable. In any case the selection of any specific value over another in any practical application should be transparently discussed with respect to the applied selection criteria.

Conclusions

We demonstrated that the well-being valuation approach *can* be another useful instrument in the (health) economist's toolbox for obtaining monetary equivalent valuations of health (v_Q). Some inherent empirical and conceptual challenges of applying this approach in this context can be addressed, especially when using large-scale longitudinal data. However, other issues, like the health state dependence of the utility of consumption, will remain a threat to the validity of estimates, warranting additional research. Concurrently, alternative approaches of estimating v_Q, like stated preference studies or methods aiming at eliciting the value of a statistical life, as recently applied by Herrera-Araujo, Hammitt, and Rheinberger (2020),[274] provide important complementary insights, despite their conceptual differences. Also given their respective strengths and limitations, methodological diversity is desired in the ongoing endeavour of measuring the monetary equivalent value of health.

The type of v_Q estimates provided in our analysis reflect average marginal health valuations (with the caveat of being entirely based on marginal changes in health-related quality of life), representative on a national level. As such, these can be applied in economic evaluations informing decision making on a societal level for publicly funded policies or interventions. Such v_Q estimates predominantly find their use by informing the cost-effectiveness threshold in the context of cost-utility analysis within health care, which aid in informing decisions on reimbursement of certain health interventions. However, estimates of the monetary value of health can also be useful in broader contexts, like cost-benefit analyses or similar approaches,[273] especially when benefits and costs of policies/interventions constitute a mix of health and non-health outcomes occurring across different sectors. Advancing methodologies aiming to estimate v_Q and providing insights into their validity can assist in informing some of the uncomfortable trade-offs that societies generally face in priority-setting both within health care but also beyond.[162]

9

General discussion

The aim of this thesis was to contribute to broadening the scope and strengthening the empirical basis of health economic evaluations. This is relevant because health economic evaluations can inform whether the costs of interventions in health and social care are proportional to the benefits they provide, thereby increasing the quality, transparency, and accountability of resource allocation decisions in health care.

While the general idea of the health economic evaluation framework is rather straightforward, its exact definition and scope as well as its empirical basis of measuring and valuing the benefit dimension is not. For example, traditionally, the scope of the outcome dimension was limited to health alone, but it has been argued that this may not be appropriate in all contexts. Using inappropriate or incomplete outcome measures may carry the risk of not maximising the value of health care interventions, as perceived and experienced by recipients of that health care. We may not be maximising what is relevant to the corresponding recipients of health care. This implies that limited health resources are not optimally allocated. This may also be the case if the monetary value of health gains is not set appropriately. Then, assessing whether the benefits of an intervention outweigh the costs associated with it, becomes troublesome. Moreover, if the scope of health economic evaluations is to be extended to measuring broader well-being, the monetary value of gains in well-being also needs to be established to make welfare improving decisions.

Therefore, further empirical, and conceptual research was (and still is) needed for facilitating and promoting the use of broader outcome measures in health economic evaluations. This final chapter summarises and discusses the overall findings of this thesis in relation to its overall aim. Limitations are highlighted alongside contributions to the literature and main implications for research and policy.

Part I: Generating weights for health and well-being instruments

The first part of this dissertation addressed conceptual and methodological questions about assessing the importance of different dimensions of well-being within several multi-dimensional well-being instruments. This is a necessary step to be able to quantify the benefits of interventions for economic evaluations.

Chapter 2 reported on a study applying an alternative estimation approach for valuing states for two broader well-being instruments, i.e., the ICECAP-A and ICECAP-O (instruments measuring capability well-being among adult and senior population, respectively). The research questions addressed in this chapter were whether it is feasible to create experienced utility tariffs for the broader ICECAP capability well-being measures with well-being data, and how these experienced utility tariffs compare to the existing decision utility tariffs. The main findings were the following: First, it proved to be possible to obtain sets of experienced utility weights with a straightforward regression-based approach. Second, although the two approaches (and what they measure) are different, the estimated weights shared a fair degree of similarity. Third, for the ICECAP-O, the largest differences in the two sets of weights were found in the enjoyment dimension, while for the ICECAP-A differences were found especially in the attachment and autonomy dimensions.

As broader well-being instruments tend to contain more quality-of-life domains, survey-based experiments for valuing well-being states become more difficult for participants.[90] This may especially be the case for older participants. Chapter 3 therefore compared two types of experiments for generating utility tariffs for the Well-being Of Older People measure (WOOP), a newly developed well-being instrument comprising nine dimensions.[26] The analysis attempted to establish whether best-worst scaling (BWS) or discrete choice experiment (DCE) is less cognitively burdensome for older participants. The results in terms of both revealed and stated cognitive burden indicated that DCE tasks were less cognitively burdensome than BWS.

Next, Chapter 4 aimed to assess the preferences of older people in the Netherlands regarding the relative importance of the nine well-being dimensions of the WOOP, a broader quality of life instrument.[26] As such, a utility tariff for the WOOP was created using a duration discrete choice experiment. The main results of the DCE indicated that 'physical health' and 'mental health' were the two most important well-being domains, followed by 'making ends meet' and 'independence'. The domains 'social contacts', 'receiving support', or 'feeling useful' were estimated to be less important.

Part II: Estimating the monetary value of health and well-being gains

The research reported on in Part II focused on estimating the monetary value of health (safety) and well-being, applying different methodologies. It also comprehensively tested the suitability of a novel approach for the monetary valuation of health gains.

Already before COVID-19, the need for investing in effective infectious disease outbreak prevention was recognized. Chapter 5 reported on a willingness-to-pay experiment to estimate the societal monetary valuation of an international integrated early warning system for infectious diseases aimed at increasing the health safety of citizens in Europe. The mean monthly willingness to pay for such a system across six European countries was €21.80 (median=€10.00). These values ranged from €8.89 (median=€3.85) in Hungary to €28.33 (median=€13.42) in Denmark. Aggregating the median values would result in a yearly contribution for the six included countries of €6.5bn. Although rigorously designed and with results behaving as expected, it needs to be acknowledged that due to the inherent limitations of willingness-to-pay-experiments, such as framing effects and hypothetical bias),[39] these valuations need to be interpreted with caution.

Chapter 6 reported on a replication of the willingness-to-pay study reported in Chapter 5 but performed in a dramatically changed context. The survey data used in chapter 5 were collected in 2018, when no immediate pandemic threat seemed to exist in Europe. This changed with the onset of the COVD-19 pandemic in early 2020. Therefore, this chapter tried to answer the question whether the valuation of an early warning system for infectious diseases changed in the wake of this pandemic, comparing data from April 2020 and April 2018. Mean monthly WTP for an early warning system for infectious diseases significantly increased, i.e., by about 50%, depending on the specified WTP scenario. The corresponding median values increased by about 30%. The results highlighted the following: First, the conducted experiment and its results were sensitive to the change in context. Second, the changes arguably were relatively moderate. Third, different scenarios in the experiment, with widely varying health benefits, led to only slightly different results, providing further evidence of insensitivity to scope in WTP experiments.

The well-being valuation approach is one of the methods that can be used for estimating the monetary value of health, or more specifically, the consumption value of health v_Q.[9,44] Chapter 7 reported on a study based on UK data applying this regression-based approach to estimate the monetary of a QALY, i.e., a year in full health, but also the monetary value of a year in full capability well-being (measured using the ICECAP-A instrument). It therefore addressed the question whether it is feasible to estimate a monetary value for both health and well-being using this approach and if so, what the relative monetary value of health gains and well-being gains. The baseline monetary estimates were £30,786 for one QALY and £66,597 for one year in full capability, with relative magnitudes varying between 1.7 and 2.6 in the conducted robustness checks. This analysis confirmed that the well-being valuation approach can produce estimates with certain face validity, as they were in line with previous estimates for health.[38] Moreover, results substantiate that capability well-being represents a broader outcome than health, as its value was estimated to be higher.

Chapter 8 aimed to assess several empirical challenges of the well-being valuation approach for estimating the monetary value of health and proposes how some of these could be addressed. In this chapter, the well-being valuation approach was applied to large-scale longitudinal data from Germany to address challenges related to the functional specification of income, the choice of health utility tariffs, and the health state dependence of consumption utility. The baseline monetary estimates the value of a QALY in Germany was €58,533, but €22,717 when instrumenting for income. Across various model specifications relating to the empirical questions under examination, the bulk of values ranged between €20,000 and €60,000 per QALY. Several recommendations for future studies using the well-being valuation approach to value health were formulated based on the results of this study.

Limitations and corresponding research avenues

While analysis-specific limitations have been discussed in each of the respective chapters, the following aims to discuss the general limitations of this thesis. Avenues for future studies are formulated as well.

The empirical research described in this thesis is often based on experimental data, eliciting stated preferences (chapters 3 to 6). The downsides of stated preference approaches, like hypothetical response bias, insensitivity to scope, or framing effects, are well known.[164] While the analysis of well-being data in the context of this thesis (chapters 1, 7, and 8) seems to be a promising alternative, the following question remains unanswered: How do the obtained preferences and valuations compare to actual decisions, behaviours and trade-offs, i.e., revealed preferences? Insights into this would strengthen the validity of many of the obtained results. The COVID-19 pandemic may actually have provided a situation where, for instance, trade-offs between different well-being domains, or specifically health and well-being, could be observed more clearly. Over the past two years, individuals have constantly been trading off their health (infection and associated risks) and dimensions that are important to well-being, like their social life. Moreover, willingness to pay for increasing in health safety are now also more visible, as individuals purchase (quality) masks or (self-)tests out of their own pocket. Such observed behaviours provide interesting opportunities for further research into the (relative) value of health and well-being.

Another limitation concerns the approach for obtaining the survey and experimental data. All data used in this dissertation (except for chapter 8) was collected through online surveys administered to (commercial) online panels. The use of such online formats and panels has been increasing and it has been shown that the quality can be comparable to mail surveys in the context of WTP experiments.[275] However, concerns do exist regarding the comparability of members of such online panels to the general population.[276] These concerns may even be stronger when aiming to include vulnerable populations like patients or older people. Given the increasing elicitation and use of patient preferences to inform policy decisions,[277] efforts for improving the representativeness in online sampling appears to be a worthwhile effort. This was recently already recognized in the context of data collections during the COVID-19 pandemic.[278]

A limitation concerning chapters 3 and 4 is that, while we attempted to assess and address cognitive burden of the choice tasks, it remains an question how to disentangle informed/preference-based choices and choices based on decision heuristics in choice experiments.[90] With the tendency towards the development of larger instruments, or experiments with a larger number of attributes in general, decision heuristics potentially play an increasingly important role. To advance health state valuation based on choice experiments, it would be beneficial to further examine the prevalence of choice heuristics in designs with varying complexity in different respondent groups. Furthermore, the implications of choice heuristics for the internal and external validity of results of choice experiments should be further investigated, and possible strategies to overcome them need to be developed.

A general limitation of the current approach to health state valuation, also applying to the WOOP utility tariffs created in chapter 4, is that only one set of utility weights based on the preferences of the general public is generally calculated and subsequently in economic evaluations of all sorts of interventions.[1] However, preferences might differ within the general population, between the general population and the target group of an intervention, often patients with a certain disease or limitation, and also between target groups of different interventions. Already in the development stage of the WOOP, it was shown that heterogeneity exists concerning what well-being means to older people.[92] Also in chapter 4 we found evidence for heterogeneity in preferences (as shown by the standard deviations in Table 9). It is a normative choice in health state valuation to apply just one utility tariff irrespective of the population of interest, usually motivated by the fact that it reflects the average preferences of the general public, who are the payers of interventions (through taxes and premiums). However, it seems worthwhile to conceptually and empirically explore estimating the cost-utility of interventions using heterogenous utility value sets, also given the emerging opportunities from advances in DCE design generation and statistical modelling.[277]

In the studies presented in chapters 5 and 6 it was not possible to overcome the major limitations of willingness-to-pay experiments in this context, which lie in hypothetical response bias and insensitivity to scope.[164] However, two lessons can be learned from these studies for WTP experiments using different risk scenarios: First, observing that many respondents anchored their responses on the first scenario, it would be recommendable to randomise the order of scenarios or to separate the scenarios in the survey. Second, given strong insensitivity to scope, which may be different for goods more tangible to respondents, it may be recommendable to use more of the respondents' time to better inform one or two basic scenarios rather than presenting them with many different scenarios.

An open, more normative, question in terms of the results the studies presented in chapters 5 and 6 remains: Seeing that the onset of the pandemic has changed the monetary valuation of health, what is the policy relevant value? The WTP obtained before the pandemic emergency, based on less informed preferences? Or the WTP obtained after respondents experienced the consequences of a specific pandemic and updated their preferences? In essence, this relates to the point raised before about using general public or target group (i.e., patient) utility tariffs for valuing the health or well-being effects of an intervention.

In terms of the findings of chapters 7 and 8, the following has to be acknowledged: The estimated monetary valuations represent just one (set of) values using one specific approach, based on specific assumptions. As outlined before, many other approaches can be used for estimating v_Q, which may lead to estimates that are conceptually different from each other and that could lead to different policy recommendations due to differences in results. For instance, our main monetary estimate of one year in *full* capability (£66,597) is an average *individual* valuation, while a recent qualitative study provided an average *societal* valuation for a year in *sufficient* capability that is substantially lower (£33,500).[134] Furthermore, the v_Q estimates reported in this dissertation relate to the consumption value of health, while for instance in the UK, threshold values are oriented on health opportunity costs (sometimes denoted k_Q).[9,279] Given this variety in perspectives and methods and the implications of findings from policy decision-making and individual patients, it is recommended that researchers should be clear and transparent about what it actually is that they are estimating. Also, it should be clear which choices have been made and why in collecting, cleaning, and analysing the data, and how these choices potentially affect their findings. Because estimates of v_Q and k_Q provide complementary insights useful for health care decision making, considering the results of chapter 8, it may be of value to provide estimates of k_Q for Germany, applying similar approaches that have been used before in the UK.[218]

A last, self-evident, limitation of this thesis is that the generated insights are context specific. For instance, the estimated monetary valuations of a QALY for the UK and Germany may not be very relevant for other countries. Similarly, the created utility tariff for the WOOP is primarily intended to be used in economic evaluations in the Netherlands. At the same time, the conceptual insights and methodological advances presented in this thesis are generalisable and applicable to the health economic evaluation framework in many more countries. Also, the presented studies may serve as blueprints for replications in other contexts and populations.

Contribution and relevance of results

On a general level, the contribution of this thesis to the literature is to have advanced the traditional health economic evaluation framework and its empirical basis in two particular ways: First, conceptual and innovative empirical work was conducted for valuing well-being states. This contributes to measuring and valuing broader quality of life gains for the purpose of economic evaluations. Second, different methodologies were applied and tested for estimating the monetary valuation of health and well-being gains. These two contributions together facilitate the incorporation of broader well-being outcomes in health economic evaluations. The following summarises the specific contributions of the chapters, alongside highlighting the policy relevance of the findings.

The study in Chapter 2 was the first to create experienced utility tariffs for broader well-being instruments. These alternative tariffs for ICECAP-O and ICECAP-A can readily be used in health economic evaluations in the UK. While these tariffs turned out to be similar to the available decision utility tariffs, this is not necessarily the case for other instruments or in other contexts. Decision utility and experienced utility provide conceptually different information, and can both be relevant to decision makers.[62] Therefore, it is recommended to use the different sets of weights next to each other and discuss the implications of differences in findings.

Previous studies on valuing well-being states exclusively applied best-worst scaling approaches,[25,33,47] while studies valuing health states predominantly used discrete choice experiments.[30] For estimating a utility tariff for WOOP, both approaches were initially considered. The finding of Chapter 3 that DCE was considered less burdensome, motivated the use of this method in Chapter 4. Although this finding was in contrast to some earlier claims,[88] it was in line with more recent research.[86] Furthermore, while our results are context specific, they highlight that conducting similar a priori experiments to inform the choice between elicitation methods is recommended.

The utility tariffs created in Chapter 4 enables the use of the WOOP in health economic evaluations in the Netherlands. Assuming that well-being preferences are somewhat overlapping in historically and culturally similar countries, the tariffs may also be used in other western, industrialized countries. That health state preferences only marginally differ between similar countries has been shown before for a different quality of life instrument.[280]

In addition, Chapter 4 provided further insights into what really matters to older people in terms of different well-being domains. These insights may be useful for decision making in terms of agenda and priority setting of policies affecting older people, both inside and outside the health care sector. For instance, the finding that mental health was perceived to be the most important well-being domain, and at least as important as physical health, implies that public investments could focus more on mental health care. It may, for example, also shed a different light on decisions to completely isolate older people living in nursing homes from their families living in the community during certain stages of the COVID-19 pandemic.

Furthermore, Chapter 4 contained some important methodological contributions. First, this study was the first to apply a DCE with duration approach for estimating a utility tariff for a broader well-being measure.[33,102] This approach allows to estimate a QALY-type tariff, anchoring the utility index on dead (0) and full well-being (1), with possible negative values representing well-being states considered worse than dead. Second, the estimated model was specified to accommodate for non-linear time preferences.[122] This resulted in an unbiased representation of the trade-off between length and quality of life inherent to such valuation studies. Consequently, the application of such models is strongly encouraged for the purpose of health state valuation using DCE with duration approaches. Third, given the overall favorable results regarding cognitive burden, despite the large descriptive system and the complexity of the tasks, it appears that the steps that were taken to reduce cognitive burden of the DCE were successful. Color coding, level overlap and separating out duration, therefore, can be recommended for future studies using DCE's in similar contexts.

The studies presented in **Chapters 5 and 6** present first monetary valuations for health safety provided by an early warning system for infectious diseases for six European countries. While point estimates from WTP studies, the predominant approach for the monetary valuation of health,[38] need to be interpreted carefully, two aspects are worth noting: First, in a more general sense, the results indicate that the majority of respondents consider such an early warning system to be of value (even before the pandemic). Second, the relative size of the monetary estimates across countries can be of value to inform public investments in pandemic prevention especially at the European level. In any case, in this post-COVID era, the need for transnational pandemic prevention has become more obvious and the studies included in this dissertation provide a starting point for informing international investment decisions in this area.

The contribution of **Chapter 7** lies in a side-by-side estimation of the monetary value of a QALY and a year in capability well-being, a novelty in the literature. The obtained estimate for one year in full capability thus represents a first approximation of the individual monetary valuation of a broader well-being measure. The results imply that if the ICECAP-A is used in economic evaluations, using a (considerably) higher threshold as compared to the QALY threshold would be appropriate, which was to be expected given the broader scope of wellbeing measures. As such, the valuation of one year in full capability, together with findings from previous related work,[134] can be used for informing the cost-effectiveness of interventions assessed using the ICECAP-A in the UK context. Furthermore, the estimated monetary value of a QALY, although not based on health opportunity costs, adds to the existing literature on the QALY threshold value in the UK.[38]

The analysis in **Chapter 8** is the most comprehensive study up to this point on applying the well-being valuation approach for estimating the monetary value of health. The implications of the results for future applications of this approach are the following: Outliers need to be addressed, different functional forms of income need to be tested, the choice of utility tariff is not neutral and needs to be well argued, and the health state dependence of consumption utility needs to be addressed.

At the same time, further research is needed on additional strategies to estimate causal impacts of health/well-being and income on life satisfaction. One potential way forward could be to establish country-specific yardsticks for the causal estimate of income on life satisfaction by using insights from randomised (social) experiments like the basic income experiment (now also launched in Germany[*]).[281] While not the primary purpose of this chapter, the estimated monetary valuations of a QALY provide a further basis for the discussion about the establishment and the height of a threshold value for Germany.

Conclusion

The necessity of performing health economic evaluations, and of extending their scope is at least partly motivated by ageing populations and rising health care costs.[2] As many western countries are only at the beginning of this demographic transition, the need and importance of helpful tools informing the allocation of scarce health care resources will likely increase. While not without flaws, cost-utility analysis appears to be the best tool for assessing the value for money of most interventions (so far).[†] At the same time, the framework for the measurement and valuation of benefits, either in terms of health or well-being, can be further developed and aligned with public preferences to further improve welfare from policy decisions. The here proposed extensions and refinements of the health economic evaluation framework are only some among many. For instance, additional instruments measuring broader well-being are currently being developed; concerns around health equity within this framework are being addressed;[14] and the inclusion of future costs, unrelated to the intervention under assessment, is being discussed.[282] On a broader scale, it may be worthwhile to investigate how cost-utility analysis can be used for informing allocation of public spending across sectors of the economy.[283]

To conclude, decisions about the allocation of scarce health care resources have to be made now and will need to be made in the future. This dissertation contributed to the growing body of literature extending the health economic evaluation framework used for informing such decisions. To maintain and improve the accessibility, affordability, and quality of health care, taking equity considerations into account, it is essential that health economists continue to operate and further refine this framework with cool heads, but warm hearts.

[*] https://www.pilotprojekt-grundeinkommen.de/english

[†] At the same time, some countries, like Germany, are still reluctant to acknowledge this.

Summary & propositions

Health care resources are scarce. Consequently, decisions must be made about how to allocate these resources. In publicly funded health care systems these decisions could, for example, be about whether a new medical intervention is included in the health benefits package, or not. Institutions responsible for making these judgements require adequate information to be able to make choices that benefit society. Health economic evaluations provide such evidence by systematically comparing the costs and benefits of interventions. Traditionally, the assessment of benefits has been exclusively focused on health effects. However, for certain health and social care services considering broader outcomes may be more appropriate, as the aim of interventions may be to improve well-being rather than only or specifically health (elderly care, social care, palliative care). In such settings, focusing on only the health effects of interventions would provide an incomplete assessment of their benefits. This, in turn, could lead to suboptimal resource allocation decisions.

Therefore, the general objective of part A of this thesis was to conduct research facilitating the use of broader outcome measures in health economic evaluations, which go beyond health and extend benefit assessment of interventions to well-being. In one study, this entailed testing the feasibility of an alternative method for weighing the dimensions of multi-dimensional outcome measures, a necessary step for quantifying benefits for health economic evaluations, with a special focus on well-being instruments. The conceptual difference of the applied method to more traditional approaches was to base the weighting on actual well-being experiences of individuals and not judgements of expected well-being experiences in hypothetical well-being states. While the resulting sets of weights were similar to existing, conventional, weights, they differed meaningfully in well-being dimensions relating to enjoyment and attachment, and in the value of small well-being decrements.

The aim of the two other studies in part A of this dissertation was informing and developing an experiment for obtaining preferences of older people regarding well-being states. In particular, these studies focused on investigating the importance of the nine dimensions of a novel well-being measure for older people. The experiment showed that mental health, physical health, being able to make ends meet (financial security) and independence were assessed to be the most important well-being dimensions by older people.

Part B of this dissertation was motivated by the fact that health economic evaluations are most likely to lead to optimal decisions if information is available on which ratio of costs and benefits is still considered acceptable, i.e., a monetary threshold value for what society is willing to pay for an additional unit of outcome. Research exists on the monetary value of health, but if the scope of the benefits under consideration is extended to well-being, new evidence is necessary about the monetary value of well-being gains. Therefore, the broader objective of part B of this thesis was to apply and refine existing methodologies to value health for estimating the societal monetary value of well-being.

One avenue of research comprised of two studies applying willingness to pay experiments in the context of a (potential) health intervention. The purpose of these experiments was to obtain a societal monetary valuation of increases in health safety provided by an integrated European early warning system for infectious diseases. A willingness to pay experiment that was fielded before the start of the COVID-19 pandemic was replicated two years later, during the pandemic. The results of these two experiments showed that most individuals would be willing to pay for the health safety benefits from such an early warning system and that the mean willingness to pay increased by about 50% after the start of the COVID-19 pandemic. The two studies also provided further insights into the sensitivity of such willingness to pay experiments to the size of the benefits considered.

A further research avenue of part B of this dissertation entailed applying and testing the well-being valuation approach for estimating the societal monetary valuation of health and well-being. This approach has not yet been used and tested for this purpose as extensively as methods like willingness to pay or discrete choice experiments. One study focused on obtaining monetary estimates for health and capability well-being based on data from the UK. This study found that a gained year in full capability well-being is valued roughly twice as high as a gained year in full health. In a second study, the well-being valuation approach was applied to a large German dataset to value health and the sensitivity of this approach to different model and variable specifications was extensively investigated. Several recommendations for future studies using the well-being approach were formulated based on the results of this study.

The overall implications of this dissertation can be summarised as follows: First, the conceptual and innovative empirical work on weighting well-being dimensions in part A of this thesis facilitates the use of well-being measures in health economic evaluations and provides guidance for future research aiming to compute such weights for well-being instruments. Secondly, the studies on estimating the monetary value of health and well-being in part B of his thesis provide useful and actionable information for health economic evaluations of interventions with a broader aim than health improvement. Moreover, the thorough investigation and discussion of the applied methods may prove valuable for the design and specification of future related studies. The research presented in this thesis contributed to the growing body of literature extending the health economic evaluation framework to include broader outcomes. Nevertheless, given the challenges arising from ageing populations and other pressures on health care budgets, the health economic evaluation framework needs to remain flexible and open to further developments in order to continue to meaningfully inform policy decisions about maintaining and improving the accessibility, affordability, and quality of health care.

Propositions

1. The adequate evaluation of health and social care services requires a scope beyond health (this dissertation).

2. The well-being valuation method is a valid additional approach for estimating the monetary value of non-market goods like health (this dissertation).

3. Utility tariffs based on decision and experienced utility are conceptually different, but both provide relevant information for reimbursement decisions (this dissertation).

4. Contingent valuation studies have their limitations when valuing complex interventions (this dissertation).

5. Discrete choice experiments with duration are able to provide anchored utility tariffs for comprehensive health or well-being instruments (this dissertation).

6. Science before statistics!

7. Open science can lead to a second credibility revolution in economic research.

8. More open and transparent peer-review processes could lead to better research.

9. The concept of opportunity costs should play a larger role in political discussions, also outside health care.

10. Large-scale policy experimentation should be used more frequently to inform important legislation.

11. Cake is better than coffee.

Nederlandse samenvatting

De middelen voor gezondheidszorg zijn schaars. Daarom moeten er beslissingen worden genomen over de beste inzet van deze middelen. In publiek gefinancierde zorgstelsels kunnen deze beslissingen bijvoorbeeld gaan over het al dan niet opnemen van een nieuwe medische interventie in het basispakket van vergoede zorg. Instellingen die verantwoordelijk zijn voor het maken van deze beslissingen hebben adequate informatie nodig om keuzes te kunnen maken die de samenleving ten goede komen. Gezondheidseconomische evaluaties leveren die informatie door de kosten en baten van interventies systematisch te vergelijken. Traditioneel ligt bij de beoordeling van de baten de nadruk op de gezondheidseffecten. Voor bepaalde interventies in de gezondheidszorg of de langdurige zorg kan het echter passender zijn om bredere uitkomsten te gebruiken, aangezien het doel van zulke interventies meer gericht is op het verbeteren van welzijn dan (alleen) gezondheid (zoals in de ouderenzorg, Maatschappelijke gezondheidszorg of palliatieve zorg). In dergelijke gevallen zou het focussen op alleen de gezondheidseffecten van interventies een onvolledig beeld van de baten van interventies opleveren. Dit zou vervolgens kunnen leiden tot suboptimale beslissingen over de inzet van schaarse zorgmiddelen.

Daarom was de algemene doelstelling van deel A van dit proefschrift om onderzoek te doen naar het gebruik van bredere uitkomstmaten in gezondheidseconomische evaluaties. Zulke uitkomstmaten meten meer dan alleen gezondheid en breiden de beoordeling van baten van interventies uit naar het meten van welzijn. In een eerste onderzoek werd de haalbaarheid getest van een alternatieve methode voor het wegen van de dimensies van multidimensionale uitkomstmaten, met speciale aandacht voor welzijnsinstrumenten. Die weging is een noodzakelijke stap in het kwantificeren van de baten van interventies in gezondheidseconomische evaluaties. Het conceptuele verschil van de hier toegepaste methode met meer traditionele benaderingen was om de weging te baseren op feitelijke welzijnservaringen van individuen en niet op beoordelingen van verwachte welzijnservaringen in hypothetische welzijnstoestanden. Hoewel de resulterende sets van gewichten vergelijkbaar waren met bestaande, conventionele gewichten, verschilden ze aanzienlijk voor de welzijnsdimensies met betrekking tot 'plezier' en 'liefde en vriendschap', en in de waardering van kleine welzijnsverminderingen. Het doel van de twee andere studies in deel A van dit proefschrift was het informeren en ontwikkelen van een experiment voor het meten van voorkeuren van ouderen met betrekking tot welzijnstoestanden. Deze studies waren met name gericht op het onderzoeken van het gewicht van de negen dimensies van een nieuwe welzijnsmaat voor ouderen. Uit het experiment bleek dat geestelijke gezondheid, lichamelijke gezondheid, rondkomen (financiële zekerheid) en zelfstandigheid door ouderen als de belangrijkste welzijnsdimensies werden beoordeeld.

Deel B van dit proefschrift had als uitgangspunt dat gezondheidseconomische evaluaties de meeste kans hebben om tot optimale beslissingen te leiden als er informatie beschikbaar is over welke verhouding tussen kosten en baten acceptabel wordt geacht, dat wil zeggen, een monetaire drempelwaarde voor wat de samenleving bereid is te betalen voor een extra eenheid van de relevante uitkomst. Er bestaat weliswaar onderzoek naar de monetaire waarde van gezondheidswinst, maar als de reikwijdte van de relevante baten wordt uitgebreid tot welzijn, is onderzoek nodig naar de monetaire waarde van welzijnswinst. Daarom was de bredere doelstelling van deel B van dit proefschrift om bestaande methodologieën om gezondheid te waarderen toe te passen en te verfijnen voor het schatten van de monetaire waarde van welzijn vanuit maatschappelijk perspectief. Een eerste onderzoekslijn bestond uit twee onderzoeken waarin experimenten werden uitgevoerd om de betalingsbereidheid voor een (potentiële) gezondheidsinterventie te meten. Het doel van deze experimenten was om een maatschappelijke monetaire waarde te schatten van een betere gezondheidsveiligheid door het opzetten van een geïntegreerd Europees systeem voor vroegtijdige waarschuwing voor infectieziekten. Een van de twee experimenten vond plaats vóór het begin van de COVID-19-pandemie. Dit experiment werd twee jaar later, tijdens de pandemie, herhaald. De resultaten van deze twee experimenten toonden aan dat de meeste mensen bereid zouden zijn te betalen voor betere gezondheidsbescherming door een dergelijk systeem voor vroegtijdige waarschuwing. De gemiddelde betalingsbereidheid nam na het begin van de COVID-19-pandemie met ongeveer 50% toe. De twee onderzoeken gaven ook meer inzicht in de gevoeligheid van de betalingsbereidheid voor de omvang van de gepresenteerde gezondheidsbaten.

Een tweede onderzoekslijn in deel B van dit proefschrift omvatte het toepassen en testen van de welzijnswaarderingsmethode voor het schatten van de maatschappelijke monetaire waarde van gezondheid en welzijn. Deze aanpak werd niet eerder zo uitgebreid gebruikt en getest voor dit doel, waarvoor meestal methoden als betalingsbereidheid en discrete keuze experimenten worden gebruikt. Eén onderzoek was gericht op het verkrijgen van monetaire waarderingen van gezondheid en welzijn op basis van data uit het VK. Uit deze studie bleek dat de waarde van een gewonnen jaar in volledig welzijn ongeveer twee keer zo hoog was als een gewonnen jaar in volledige gezondheid. In een tweede studie werd de welzijnswaarderingsmethode toegepast om gezondheid te waarderen. Hierbij werd gebruik gemaakt van een grote Duitse dataset en werd de gevoeligheid van de uitkomsten voor verschillen in de gebruikte specificaties van modellen en variabelen uitgebreid onderzocht. Op basis van de resultaten van dit onderzoek zijn verschillende aanbevelingen geformuleerd voor toekomstig onderzoek waarin de welzijnswaarderingsmethode wordt toegepast.

De algemene implicaties van de bevindingen in dit proefschrift kunnen als volgt worden samengevat: Ten eerste, vergemakkelijkt het conceptuele en innovatieve empirische werk over het wegen van welzijnsdimensies, gepresenteerd in deel A van dit proefschrift, het gebruik van welzijnsmaten in gezondheidseconomische evaluaties. Daarnaast biedt het een leidraad voor toekomstig onderzoek gericht op het berekenen van dergelijke gewichten voor welzijnsinstrumenten. Ten tweede, leveren de studies in deel B van zijn proefschrift, over het schatten van de monetaire waarde van gezondheid en welzijn, bruikbare en toepasbare informatie voor gezondheidseconomische evaluaties van interventies die bredere uitkomsten hebben dan alleen gezondheidsverbetering. Bovendien kan het grondige onderzoek naar en de bespreking van de toegepaste methoden waardevol zijn voor het ontwerp en de specificatie van toekomstige studies op dit terrein.

Het onderzoek dat in dit proefschrift werd gepresenteerd draagt bij aan de wetenschappelijke kennis over het uitbreiden van het raamwerk van gezondheidseconomische evaluaties om bredere uitkomstmaten te kunnen includeren. Desalniettemin moet dit raamwerk, gezien de uitdagingen die voortvloeien uit de vergrijzing van de bevolking en de bredere druk op zorgbudgetten, open blijven staan voor verdere ontwikkelingen. Dit is nodig om beleidsbeslissingen op zinvolle wijze te blijven informeren en uiteindelijk de toegankelijkheid, betaalbaarheid en kwaliteit van de zorg te bewaken en te verbeteren.

PhD portfolio

PhD Candidate	Sebastian F.W. Himmler
Promotor	Werner B.F. Brouwer
Co-promotor	Job N.A. van Exel
PhD Period	September 2017-August 2022

Training activities	Year	ECTS
European Training Network courses		
- Epidemiology and Economics	2017	5.0
- Microeconometrics	2017	5.0
- Working with SAS	2018	5.0
- Experimental Design	2018	5.0
- Survey Design and effectiveness research	2018	5.0
- Measuring Quality of Care Using Administrative Data	2018	5.0
- Economic Evaluation and Quality of Care	2018	5.0
- Defining and Measuring Patient Satisfaction: Current Issues and concepts	2018	5.0
- Performance Measurement and Multilevel Modelling	2018	5.0
- Writing Skills & Intellectual Property Rights & Research Integrity	2017	
- Project Management & Time- and Self-Management Skills	2017	
- Communication and Presentation Skills	2018	
- Leadership- and Team Building Training	2019	
- Funding Opportunities & Drafting a Research Proposal	2019	
- Interpersonal and Networking Skills	2019	
Further courses & workshops		
- Making an academic poster that stands out	2018	1.0
- Self-presentation: confidence, focus, persuasion	2018	2.5
- Choice Modelling and Stated Preference Design course	2019	1.0
- Analytic storytelling	2020	2.5
- Digital research methods for textual data	2020	2.5
- Winter School in Data Analytics and Machine Learning	2021	
- Data visualization – the art/skill cocktail	2021	

Teaching activities	
Lectures & Practicals	
- Kwantitatief Leeronderzoek (KLO)	2019
- Global Health Economics (auxiliary practical staff)	2019
- Global Health Economics (practicals + 0.5 lectures)	2020
- Measurement of patient preferences using Discrete Choice Experiments (auxiliary practical staff)	2020
- Global Health Economics (practicals + 1.0 lectures)	2021
Supervision	
- Master thesis (2/2 students graduated)	2019
- Master thesis (2/2 students graduated)	2020
- Master thesis (2/2 students graduated)	2021

Conferences with presentation	
- EuHEA (Maastricht)	2018
- EuHEA PhD Conference (Catania)	2018
- iHEA (Basel)	2019
- NHESG (Rejkjavik)	2019
- EuHEA (Oslo – online)	2020
- DGGOE (Nürnberg – online)	2021

Other activities	
- Reviewer for *Economics & Human Biology* (1)	2019
- Reviewer for *The European Journal of Health Economics* (1)	2021
- Reviewer for *Value in Health* (1)	2021
- Reviewer for *Health Economics* (1)	2021

List of publications

Included in this dissertation

Himmler, S., van Exel, J., Brouwer, W., 2020. Happy with Your Capabilities? Valuing ICECAP-O and ICECAP-A States Based on Experienced Utility Using Subjective Well-Being Data. *Med. Decis. Mak*. 40, 498–510. https://doi.org/10.1177/0272989X20923015

Himmler, S., van Exel, J., Perry-Duxbury, M., Brouwer, W., 2020. Willingness to pay for an early warning system for infectious diseases. *Eur. J. Heal. Econ*. 21, 763–773. https://doi.org/10.1007/s10198-020-01171-2

Himmler, S., van Exel, J., Brouwer, W., 2020. Estimating the monetary value of health and capability well-being applying the well-being valuation approach. *Eur. J. Heal. Econ*. 21, 1235–1244. https://doi.org/10.1007/s10198-020-01231-7

Himmler, S., Soekhai, V., van Exel, J., Brouwer, W., 2021. What works better for preference elicitation among older people? Cognitive burden of discrete choice experiment and case 2 best-worst scaling in an online setting. *J. Choice Model.* 38, 100265. https://doi.org/10.1016/j.jocm.2020.100265

Himmler, S., Stöckel, J., van Exel, J., Brouwer, W., 2021. The value of health—Empirical issues when estimating the monetary value of a quality-adjusted life year based on well-being data. *Health Econ.* 30 (8), 1849-1870. https://doi.org/10.1002/hec.4279

Himmler, S., van Exel, J., Brouwer, W., 2022. Did the COVID-19 pandemic change the willingness to pay for an early warning system for infectious diseases in Europe? *Eur. J. Heal. Econ*. 23, 81-94.

Himmler, S., Jonker, M., van Krugten, F., van Exel, J., Brouwer, W., 2022. Estimating an anchored utility tariff for the well-being of older people measure (WOOP) for the Netherlands. *Soc. Sci. Med*. 301, 114901.
https://doi.org/10.1016/j.socscimed.2022.114901

Not included in this dissertation

Phillips, E., Himmler, S., Schreyögg, J., 2021. Preferences for e-Mental Health Interventions in Germany: A Discrete Choice Experiment. *Value Heal.* 24, 421–430. https://doi.org/10.1016/j.jval.2020.09.018

Enzing, J.J., Himmler, S., Knies, S., Brouwer, W.B.F., 2021. Do Profit Margins of Pharmaceuticals Influence Reimbursement Decisions? A Discrete Choice Experiment Among Dutch Healthcare Decision Makers. *Value Heal.* 25 (2), 222-229. https://doi.org/10.1016/j.jval.2021.08.007

Phillips, E., Himmler, S., Schreyögg, J., 2022. Preferences of psychotherapists for blended care in Germany: a discrete choice experiment. *BMC Psychiatry*. 22, 1–12. https://doi.org/10.1186/s12888-022-03765-x

List of publications

Acknowledgements

Looking back over the 4-5 years that it took me to put together this dissertation, I realized (among other things) that this accomplishment would not have been possible (or at least a lot less fun) without the support from so many people.

I write this on my way towards my, let's call it small 'sabbatical' after finishing my time at ESHPM. As such, I will keep it short and efficient, which is also fitting as these features are also among my most prominent characteristics. If anyone who reads the following and does not recognise their contribution to this thesis and feels cheated out, I therefore apologise for my brevity.

First of all, I would like to thank my promotors Job van Exel and Werner Brouwer. It is difficult to assess how large your contribution to this thesis is. While I am generally hesitant with such formulations when writing papers, I would say it cannot be overestimated. After our first meeting together, back in September 2017, I already had the feeling that this is going to work out great. Not just because I was awed by their depth of knowledge and their numerous ideas, but because I, rightfully, felt that we would be able to get along well on a personal level. While there naturally were ups and downs in my whole PhD trajectory, with some minor setbacks or disappointments, it was always nice to know that we would soon have a pleasant, constructive, and often also partly entertaining, meeting, which would put the disappointments into perspective. I would also like to thank them for allowing me to explore many different research ideas, while also giving me the freedom to decide about which ideas to follow-up on. Even in projects where you were less involved with, your feedback was always helpful and highly appreciated. What I also highly appreciated was the general level of pragmatism regarding research but also especially organisational matters, which fit well with my own inclination.

A second group of people I would like to acknowledge for their role in the dissertation are my (mostly 8th floor) colleagues at ESHPM. From the start, I always felt welcomed, and I always enjoyed the pleasant working environment. During the many talks in the hallways or the lunch breaks, I got to know many great people. Also, a special thanks to my PhD peers at the HE (and HE+) section. It was great to have a close group of people, who were on similar PhD timelines, with whom you could talk about successes, frustrations, or basically about anything, over lunch, cake, or a beer. In retrospect, I would say we did have a lot of fun! This definitely contributed to the fact that I never really dreaded going back to work (even on Mondays). While we were mostly working on our own projects, it felt like we were one large "support group".

Besides my support group at ESHPM, I would also like to acknowledge the role of the people involved in the ETN programme I was part of. Being able to regularly meet and discuss health economics or the experiences being a PhD with such a pleasant, international, and diverse group of people was insightful and inspiring.

Generally, I would also like to thank my family and friends for their support over the previous years. Occasionally discussing my research with you was not easy, but often

helped me to see things from different angles. At the same time, you contributed to keep me down to earth, not only as a researcher but also as a person.

Second last, but foremost, I would like to specifically thank my parents. Besides for everything else I am and could be thanks to them, I would like to note my appreciation for their patience regarding my studies. While mentioned occasionally, the "when will you finally be finished studying" also during my PhD trajectory, was never too on the nose.

I will close the acknowledgements with a last person to mention, paraphrasing lyrics from a well-known American rapper. The following quote should be interpreted in terms of self-appreciation in the humblest way possible:

"Last but not least, I would like to thank me. I wanna thank me for always believing in me. I wanna thank me for doing all this hard work, I wanna thank me for having no days off [not entirely true], I wanna thank me for never quitting."

About the author

Sebastian F.W. Himmler is a health economist, who was awarded a PhD in the field from the Erasmus School of Health Policy & Management, Rotterdam, in 2021. He was born on the 2nd of December 1988 in Sulzbach-Rosenberg, Bavaria, Germany. He obtained a bachelor (2013) and master's degree (2016) in health economics from the University of Bayreuth (Germany), with a 1-year stay as visiting scholar the University of North Carolina at Chapel Hill, NC, USA. Before starting his PhD trajectory, he worked as a research associate in the Health Economics section of the WifOR GmbH, Frankfurt, Germany, an economic research institute and think-tank.

In 2017, Sebastian successfully applied for a Marie-Curie fellowship awarded by the European Commission. The fellowship was part of a European PhD training network entitled "Improving Quality of Care in Europe (IQCE)" and funded his position as PhD candidate at the Erasmus School of Health Policy & Management in Rotterdam. The PhD trajectory of the European training network consisted of a course oriented first 1,5 years (see PhD portfolio) with a research focus in the remainder.

In his doctorate, he was applying economic and statistical tools, including applied econometrics, health economic modelling, and discrete choice experiments, to conduct research on topics surrounding tools to inform the efficient allocation of health care resources. This especially includes eliciting preferences on the importance of different (health-related) quality of life dimensions and obtaining societal monetary valuations of health and well-being gains. Further research interests included the measurement of well-being, fair pricing of pharmaceuticals, and patient and provider preferences in mental health care.

After a year as a post-doc researcher at the Erasmus School of Health Policy & Management, working on topics relating to health and well-being in times of the pandemic, Sebastian will be starting a post-doc trajectory at the Technical University of Munich in September 2022.

References

1. Drummond M, Sculpher M, Claxton K, Stoddart G, Torrance G. *Methods for the Economic Evaluation of Health Care Programmes*. 4th ed. Oxford University Press; 2015.
2. de Meijer C, Wouterse B, Polder J, Koopmanschap M. The effect of population aging on health expenditure growth: A critical review. *Eur J Ageing*. 2013;10(4):353-361. doi:10.1007/s10433-013-0280-x
3. Owens D, Siegel JE, Sculpher M, Salomon JA. Designing a Cost-Effectiveness Analysis. In: Neumann PJ, Ganiats TG, Russell LB, Sanders GD, Siegel JE, eds. *Cost-Effectiveness in Health and Medicine*. 2nd ed. Oxford University Press; 2016.
4. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20(10):1727-1736. doi:10.1007/s11136-011-9903-x
5. Brazier JE, Roberts J. The Estimation of a Preference-Based Measure of Health From the SF-12. *Med Care*. 2004;42(9):851-859.
6. Feeny D, Krahn M, Prosser LA, Salomon JA. Valuing Health Outcomes. In: Neumann PJ, Ganiats TG, Russell LB, Sanders GD, Siegel JE, eds. *Cost-Effectiveness in Health and Medicine*. 2nd ed. Oxford University Press; 2016.
7. Versteegh M, M. Vermeulen K, M. A. A. Evers S, de Wit GA, Prenger R, A. Stolk E. Dutch Tariff for the Five-Level Version of EQ-5D. *Value Heal*. 2016;19(4):343-352. doi:10.1016/j.jval.2016.01.003
8. Gold MR, Siegel JE, Russell LB, Weinstein MC. *Cost-Effectiveness in Health and Medicine*. Oxford University Press; 1996.
9. Brouwer W, van Baal P, van Exel J, Versteegh M. When is it too expensive? Cost-effectiveness thresholds and health care decision-making. *Eur J Heal Econ*. 2019;20(2):175-180. doi:10.1007/s10198-018-1000-4
10. Williams A. Qalys and ethics: A health economist's perspective. *Soc Sci Med*. 1996;43(12):1795-1804. doi:10.1016/S0277-9536(96)00082-2
11. Williams A. Cost-effectiveness analysis : is it ethical ? *J Med Ethics*. 1992;(18):7-11.
12. Pinkerton SD, Johnson-Masotti AP, Derse A, Layde PM. Ethical issues in cost-effectiveness analysis. *Eval Program Plann*. 2002;25(1):71-83. doi:10.1016/S0149-7189(01)00050-7
13. Brock D, Daniels N, Neumann PJ, Siegel JE. Ethical and Distributive Considerations. In: Neumann PJ, Ganiats TG, Russell LB, Sanders GD, Siegel JE, eds. *Cost-Effectiveness in Health and Medicine*. 2nd ed. Oxford University Press; 2016.
14. Reckers-Droog V. *Giving Weight to Equity: Improving Priority Setting in Health Care*.; 2021.
15. Ward T, Mujica-Mota RE, Spencer AE, Medina-Lara A. Incorporating Equity Concerns in Cost-Effectiveness Analyses: A Systematic Literature Review. *Pharmacoeconomics*. 2022;40(1):45-64. doi:10.1007/s40273-021-01094-7
16. Basu A. Estimating Costs and Valuations of Non-Health Benefits in Cost-Effectiveness Analysis. In: Neumann PJ, Ganiats TG, Russell LB, Sanders GD, Siegel JE, eds. *Cost-Effectiveness in Health and Medicine*. 2nd ed. Oxford University Press; 2016.
17. van Baal P, Meltzer D, Brouwer W. Future Costs, Fixed Healthcare Budgets, and the Decision Rules of Cost-Effectiveness Analysis. *Health Econ*. 2016;25(2):237-248. doi:10.1002/hec.3138
18. Coast J. Is economic evaluation in touch with society's health values? *BMJ*. 2004;329(7476):1233-1236. doi:10.1136/bmj.329.7476.1233
19. Coast J. Strategies for the economic evaluation of end-of-life care: Making a case for the capability approach. *Expert Rev Pharmacoeconomics Outcomes Res*. 2014;14(4):473-482. doi:10.1586/14737167.2014.914436
20. Milte CM, Walker R, Luszcz MA, Lancsar E, Kaambwa B, Ratcliffe J. How important is health status in defining quality of life for older people? An exploratory study of the views of older South Australians. *Appl Health Econ Health Policy*. 2014;12(1):73-84. doi:10.1007/s40258-013-0068-3
21. Payne K, McAllister M, Davies LM. Valuing the economic benefits of complex interventions: When maximising health is not sufficient. *Health Econ*. 2013;22(3):258-271. doi:10.1002/hec.2795
22. Weatherly H, Drummond M, Claxton K, et al. Methods for assessing the cost-effectiveness of public health interventions: Key challenges and recommendations. *Health Policy (New York)*. 2009;93(2-3):85-92. doi:10.1016/j.healthpol.2009.07.012
23. Grewal I, Lewis J, Flynn T, Brown J, Bond J, Coast J. Developing attributes for a generic quality of life measure for older people: Preferences or capabilities? *Soc Sci Med*. 2006;62(8):1891-1901. doi:10.1016/j.socscimed.2005.08.023
24. Al-Janabi H, N Flynn T, Coast J. Development of a self-report measure of capability wellbeing for adults: the ICECAP-A. *Qual Life Res*. 2012;21(1):167-176. doi:10.1007/s11136-011-9927-2
25. Netten A, Burge P, Malley J, et al. Outcomes of social care for adults: developing a preference-weighted measure. *Health Technol Assess (Rockv)*. 2012;16(16):1-166. doi:10.3310/hta16160
26. Hackert MQN, van Exel J, Brouwer WBF. Well-being of Older People (WOOP):Quantitative Validation of a New Outcome Measure for Use in Economic Evaluations. *Soc Sci Med*. 2020;259(April):113109.

doi:10.1016/j.socscimed.2020.113109

27. Dolan P. Whose Preferences Count ? *Med Decis Mak*. Published online 1999:482-486.

28. Cubi-Molla P, Shah K, Burström K. Experience-Based Values: A Framework for Classifying Different Types of Experience in Health Valuation Research. *Patient*. 2018;11(3):253-270. doi:10.1007/s40271-017-0292-2

29. Soekhai V, Whichello C, Levitan B, et al. Methods for exploring and eliciting patient preferences in the medical product lifecycle: a literature review. *Drug Discov Today*. 2019;24(7):1324-1331. doi:10.1016/j.drudis.2019.05.001

30. Mulhern B, Norman R, Street DJ, Viney R. One Method, Many Methodological Choices: A Structured Review of Discrete-Choice Experiments for Health State Valuation. *Pharmacoeconomics*. Published online 2018. doi:10.1007/s40273-018-0714-6

31. Soekhai V, de Bekker-Grob EW, Ellis A, Vass CM. Discrete Choice Experiments in Health Economics: Past, Present and Future. *Pharmacoeconomics*. 2018;(0123456789). doi:10.1007/s40273-018-0734-2

32. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health Econ*. 2018;27(1):7-22. doi:10.1002/hec.3564

33. Coast J, Flynn TN, Natarajan L, et al. Valuing the ICECAP capability index for older people. *Soc Sci Med*. 2008;67(5):874-882. doi:10.1016/j.socscimed.2008.05.015

34. Cameron D, Ubels J, Norström F. On what basis are medical cost-effectiveness thresholds set? Clashing opinions and an absence of data: a systematic review. *Glob Health Action*. 2018;11(1). doi:10.1080/16549716.2018.1447828

35. Cleemput I, Neyt M, Thiry N, De Laet C, Leys M. Using threshold values for cost per quality-adjusted life-year gained in healthcare decisions. *Int J Technol Assess Health Care*. 2011;27(1):71-76. doi:10.1017/S0266462310001194

36. Culyer A, McCabe C, Briggs A, et al. Searching for a threshold, not setting one: the role of the National Institute for Health and Clinical Excellence. *J Health Serv Res Policy*. 2007;12(1):56-58. doi:10.1258/135581907779497567

37. Baker R, Donaldson C, Mason H, Jones-Lee M. Willingness to Pay for Health. In: *Encyclopedia of Health Economics*. Elsevier; 2014:495-501. doi:10.1016/B978-0-12-375678-7.00503-4

38. Ryen L, Svensson M. The Willingness to Pay for a Quality Adjusted Life Year: A Review of the Empirical Literature. *Health Econ*. 2015;24(10):1289-1301. doi:10.1002/hec.3085

39. Kling CL, Phaneuf DJ, Zhao J. From Exxon to BP: Has some number become better than no number. *J Econ Perspect*. 2012;26(4):3-26. doi:10.1257/jep.26.4.3

40. Ahlert M, Breyer F, Schwettmann L. How you ask is what you get: Framing effects in willingness-to-pay for a QALY. *Soc Sci Med*. 2016;150:40-48. doi:10.1016/J.SOCSCIMED.2015.11.055

41. Bobinac A, van Exel NJA, Rutten FFH, Brouwer WBF. GET MORE, PAY MORE? An elaborate test of construct validity of willingness to pay per QALY estimates obtained through contingent valuation. *J Health Econ*. 2012;31(1):158-168. doi:10.1016/j.jhealeco.2011.09.004

42. Mason H, Jones-Lee M, Donaldson C. Modelling the monetary value of a QALY: a new approach based on UK data. *Health Econ*. 2009;18(8):933-950. doi:10.1002/hec.1416

43. Phelps CE. A New Method to Determine the Optimal Willingness to Pay in Cost-Effectiveness Analysis. *Value Heal*. 2019;22(7):785-791. doi:10.1016/j.jval.2019.03.003

44. Huang L, Frijters P, Dalziel K, Clarke P. Life satisfaction , QALYs , and the monetary value of health. *Soc Sci Med*. 2018;211(June):131-136. doi:10.1016/j.socscimed.2018.06.009

45. Mathes T, Jacobs E, Morfeld J-C, Pieper D. Methods of international health technology assessment agencies for economic evaluations- a comparative analysis. *BMC Health Serv Res*. 2013;13(1):371. doi:10.1186/1472-6963-13-371

46. Neumann PJ, Sanders GD, Russell LB, Siegel JE, Ganiats TG. *Cost Effectiveness in Health and Medicine*. Oxford University Press; 2016.

47. Flynn TN, Huynh E, Peters TJ, et al. Scoring the Icecap-a Capability Instrument. Estimation of a UK General Population Tariff. *Health Econ*. 2015;24(3):258-269. doi:10.1002/hec.3014

48. Makai P, Koopmanschap MA, Brouwer WB, Nieboer AA. A validation of the ICECAP-O in a population of post-hospitalized older people in the Netherlands. *Health Qual Life Outcomes*. 2013;11(1):57. doi:10.1186/1477-7525-11-57

49. Hackert MQN, Exel J van, Brouwer WBF. Valid Outcome Measures in Care for Older People: Comparing the ASCOT and the ICECAP-O. *Value Heal*. 2017;20(7):936-944. doi:10.1016/j.jval.2017.03.012

50. Goranitis I, Coast J, Day E, et al. Measuring Health and Broader Well-Being Benefits in the Context of Opiate Dependence: The Psychometric Performance of the ICECAP-A and the EQ-5D-5L. *Value Heal*. 2016;19(6):820-828. doi:10.1016/j.jval.2016.04.010

51. Keeley T, Al-Janabi H, Nicholls E, Foster NE, Jowett S, Coast J. A longitudinal assessment of the responsiveness of the ICECAP-A in a randomised controlled trial of a knee pain intervention. *Qual Life*

Res. 2015;24(10):2319-2331. doi:10.1007/s11136-015-0980-0

52. Al-Janabi H, Peters TJ, Brazier J, et al. An investigation of the construct validity of the ICECAP-A capability measure. *Qual Life Res*. 2013;22(7):1831-1840. doi:10.1007/s11136-012-0293-5
53. Hackert MQN, van Exel J, Brouwer WBF. Does the ICECAP-O cover the physical, mental and social functioning of older people in the UK? *Qual Life Res*. Published online November 11, 2018. doi:10.1007/s11136-018-2042-x
54. Flynn TN, Huynh E, Peters TJ, et al. Scoring the ICECAP-A capability instrument. Estimation of a UK general population tariff. *Heal Econ (United Kingdom)*. 2015;24(3):258-269. doi:10.1002/hec.3014
55. Brazier J, Ratcliffe J, Saloman J, Tsuchiya A. *Measuring and Valuing Health Benefits for Economic Evaluation*. Second edi. Oxford University Press; 2016.
56. Brazier J, Rowen D, Karimi M, Peasgood T, Tsuchiya A, Ratcliffe J. Experience-based utility and own health state valuation for a health state classification system: why and how to do it. *Eur J Heal Econ*. 2017;(0123456789):1-11. doi:10.1007/s10198-017-0931-5
57. Mann R, Brazier J, Tsuchiya A. A comparison of patient and general population weightings of EQ-5D dimensions. *Health Econ*. 2009;18(3):363-372. doi:10.1002/hec.1362
58. Little MHR, Reitmeir P, Peters A, Leidl R. The impact of differences between patient and general population EQ-5D-3l values on the mean tariff scores of different patient groups. *Value Heal*. 2014;17(4):364-371. doi:10.1016/j.jval.2014.02.002
59. Kahneman D, Sugden R. Experienced utility as a standard of policy evaluation. *Environ Resour Econ*. 2005;32(1):161-181. doi:10.1007/s10640-005-6032-4
60. Powdthavee N. What happens to people before and after disability? Focusing effects, lead effects, and adaptation in different areas of life. *Soc Sci Med*. 2009;69(12):1834-1844. doi:10.1016/j.socscimed.2009.09.023
61. Hond A De, Bakx P, Versteegh M. Can time heal all wounds ? An empirical assessment of adaptation to functional limitations. *Soc Sci Med*. 2018;222(June 2018):180-187. doi:S0277953618307111
62. Versteegh MM, Brouwer WBF. Patient and general public preferences for health states: A call to reconsider current guidelines. *Soc Sci Med*. 2016;165:66-74. doi:10.1016/j.socscimed.2016.07.043
63. Dolan P, Kahneman D. Interpretations of utility and their implications for the valuation of health*. *Econ J*. 2007;118(525):215-234. doi:10.1111/j.1468-0297.2007.02110.x
64. Dolan P, Lee H, King D, Metcalfe R. Valuing health directly. *BMJ*. 2009;339(jul20 3):b2577-b2577. doi:10.1136/bmj.b2577
65. Ferrer-i-Carbonell A, Frijters P. How Important is Methodology for the estimates of the determinants of Happiness?*. *Econ J*. 2004;114(497):641-659. doi:10.1111/j.1468-0297.2004.00235.x
66. Dolan P, Lee H, Peasgood T. Losing sight of the wood for the trees: some issues in describing and valuing health, and another possible approach. *Pharmacoeconomics*. 2012;30(11):1035-1049. doi:10.2165/11593040-000000000-00000
67. Dolan P, Metcalfe R. Valuing health: A brief report on subjective well-being versus preferences. *Med Decis Mak*. 2012;32(4):578-582. doi:10.1177/0272989X11435173
68. Mukuria C, Brazier J. Valuing the EQ-5D and the SF-6D health states using subjective well-being: A secondary analysis of patient data. *Soc Sci Med*. 2013;77(1):97-105. doi:10.1016/j.socscimed.2012.11.012
69. Cantril H. *The Pattern of Human Concerns*. Rutgers University Press; 1965.
70. Diener E, Emmons RA, Larsen RJ, Griffin S. The Satisfaction With Life Scale. *J Pers Assess*. 1985;49(1):71-75. doi:10.1207/s15327752jpa4901_13
71. OECD. *OECD Guidelines on Measuring Subjective Well-Being*. OECD Publishing; 2013. doi:10.1787/9789264191655-en
72. Helliwell JF, Barrington-Leigh C, Harris A, Huang H. International Evidence on the Social Context of Well-Being. In: *International Differences in Well-Being*. Oxford University Press; 2010:291-327. doi:10.1093/acprof:oso/9780199732739.003.0010
73. Kapteyn A, Lee J, Tassot C, Vonkova H, Zamarro G. *Dimensions of Subjective Well-Being*. Vol 123. Springer Netherlands; 2015. doi:10.1007/s11205-014-0753-0
74. Dolan P, Peasgood T, White M. Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. *J Econ Psychol*. 2008;29(1):94-122. doi:10.1016/J.JOEP.2007.09.001
75. Sprangers MA, Schwartz CE. Integrating response shift into health-related quality of life research: a theoretical model. *Soc Sci Med*. 1999;48(11):1507-1515. doi:S0277953699000453 [pii]
76. Dolan P, Lee H, Peasgood T. Losing Sight of the Wood for the Trees. *Pharmacoeconomics*. 2012;30(11):1035-1049. doi:10.2165/11593040-000000000-00000
77. Diener, Suh EM, Lucas RE, Smith HL. Subjective well-being: Three decades of progress. *Psychol Bull*. 1999;125(2):276-302.
78. Kahneman D, Wakker PP, Sarin R. Back to Bentham? Explorations of Experienced Utility. *Q J Econ*.

1997;112(2):375-406. doi:10.1162/003355397555235

79. Veenhoven R. Healthy happiness: effects of happiness on physical health and the consequences for preventive health care. *J Happiness Stud*. 2008;9(3):449-469. doi:10.1007/s10902-006-9042-1

80. Cookson R. QALYs and the capability approach. *Health Econ*. 2005;14(8):817-829. doi:10.1002/hec.975

81. Makai P, Brouwer WBF, Koopmanschap MA, Stolk EA, Nieboer AP. Quality of life instruments for economic evaluations in health and social care for older people: A systematic review. *Soc Sci Med*. 2014;102:83-93. doi:10.1016/j.socscimed.2013.11.050

82. Ryan M, Gerard K, Amaya-Amaya M, eds. *Using Discrete Choice Experiments to Value Health and Health Care*. Vol 11. Springer Netherlands; 2008. doi:10.1007/978-1-4020-5753-3

83. Flynn TN, Marley AAJ. Best-worst scaling: theory and methods. In: *Handbook of Choice Modelling*. Edward Elgar Publishing; 2014:178-201. doi:10.4337/9781781003152.00014

84. Cheung KL, Wijnen BFM, Hollin IL, et al. Using Best–Worst Scaling to Investigate Preferences in Health Care. *Pharmacoeconomics*. 2016;34(12):1195-1209. doi:10.1007/s40273-016-0429-5

85. Krucien N, Watson V, Ryan M. Is Best–Worst Scaling Suitable for Health State Valuation? A Comparison with Discrete Choice Experiments. *Heal Econ (United Kingdom)*. 2017;26(12):e1-e16. doi:10.1002/hec.3459

86. Whitty JA, Oliveira Gonçalves AS. A Systematic Review Comparing the Acceptability, Validity and Concordance of Discrete Choice Experiments and Best–Worst Scaling for Eliciting Preferences in Healthcare. *Patient*. 2018;11(3):301-317. doi:10.1007/s40271-017-0288-y

87. Louviere J. *Random Utility Theory-Based Stated Preference Elicitation Methods: Applications in Health Economics with Special Reference to Combining Sources of Preference Data*.; 2004.

88. Flynn TN. Valuing citizen and patient preferences in health: Recent developments in three types of best-worst scaling. *Expert Rev Pharmacoeconomics Outcomes Res*. 2010;10(3):259-267. doi:10.1586/erp.10.29

89. Potoglou D, Burge P, Flynn T, et al. Best-worst scaling vs. discrete choice experiments: An empirical comparison using social care data. *Soc Sci Med*. 2011;72(10):1717-1727. doi:10.1016/j.socscimed.2011.03.027

90. Jonker MF, Donkers B, de Bekker-Grob E, Stolk EA. Attribute level overlap (and color coding) can reduce task complexity, improve choice consistency, and decrease the dropout rate in discrete choice experiments. *Heal Econ (United Kingdom)*. 2019;28(3):350-363. doi:10.1002/hec.3846

91. Milte R, Ratcliffe J, Chen G, Lancsar E, Miller M, Crotty M. Cognitive overload? An exploration of the potential impact of cognitive functioning in discrete choice experiments with older people in health care. *Value Heal*. 2014;17(5):655-659. doi:10.1016/j.jval.2014.05.005

92. Hackert MQN, Brouwer WBF, Hoefman RJ, van Exel J. Views of older people in the Netherlands on wellbeing: A Q-methodology study. *Soc Sci Med*. 2019;240(December 2018):112535. doi:10.1016/j.socscimed.2019.112535

93. Jonker MF, Donkers B, de Bekker-Grob EW, Stolk EA. Effect of Level Overlap and Color Coding on Attribute Non-Attendance in Discrete Choice Experiments. *Value Heal*. 2018;21(7):767-771. doi:10.1016/j.jval.2017.10.002

94. Maddala T, Phillips KA, Johnson FR. An experiment on simplifying conjoint analysis designs for measuring preferences. *Health Econ*. 2003;12(12):1035-1047. doi:10.1002/hec.798

95. Palan S, Schitter C. Prolific.ac—A subject pool for online experiments. *J Behav Exp Financ*. 2018;17:22-27. doi:10.1016/j.jbef.2017.12.004

96. Huynh E, Coast J, Rose J, Kinghorn P, Flynn T. Values for the ICECAP-Supportive Care Measure (ICECAP-SCM) for use in economic evaluation at end of life. *Soc Sci Med*. 2017;189:114-128. doi:10.1016/j.socscimed.2017.07.012

97. Arendts G, Jan S, Beck MJ, Howard K. Preferences for the emergency department or alternatives for older people in aged care: A discrete choice experiment. *Age Ageing*. 2017;46(1):124-129. doi:10.1093/ageing/afw163

98. Franco MR, Howard K, Sherrington C, Rose J, Ferreira PH, Ferreira ML. Smallest worthwhile effect of exercise programs to prevent falls among older people: estimates from benefit-harm trade-off and discrete choice methods. *Age Ageing*. 2016;45(6):806-812. doi:10.1093/ageing/afw110

99. Jonker MF, Attema AE, Donkers B, Stolk EA, Versteegh MM. Are Health State Valuations from the General Public Biased? A Test of Health State Reference Dependency Using Self-assessed Health and an Efficient Discrete Choice Experiment. *Heal Econ (United Kingdom)*. 2017;26(12):1534-1547. doi:10.1002/hec.3445

100. Yao RT, Scarpa R, Rose JM, Turner JA. Experimental Design Criteria and Their Behavioural Efficiency: An Evaluation in the Field. *Environ Resour Econ*. 2015;62(3):433-455. doi:10.1007/s10640-014-9823-7

101. de Winter J, Dodou D. Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon (Addendum added October 2012). *Pract Assessment, Res Eval*. 2010;15(11). doi:https://doi.org/10.7275/bj1p-ts64

102. Netten A, Burge P, Malley J, et al. Outcomes of social care for adults: developing a preference-

weighted measure. *Health Technol Assess (Rockv)*. 2012;16(16). doi:10.3310/hta16160

103. Whitty JA, Walker R, Golenko X, Ratcliffe J. A think aloud study comparing the validity and acceptability of discrete choice and best worst scaling methods. *PLoS One*. 2014;9(4). doi:10.1371/journal.pone.0090635

104. King MT, Viney R, Simon Pickard A, et al. Australian Utility Weights for the EORTC QLU-C10D, a Multi-Attribute Utility Instrument Derived from the Cancer-Specific Quality of Life Questionnaire, EORTC QLQ-C30. *Pharmacoeconomics*. 2018;36(2):225-238. doi:10.1007/s40273-017-0582-5

105. Mulhern B, Norman R, De Abreu Lourenco R, Malley J, Street D, Viney R. Investigating the relative value of health and social care related quality of life using a discrete choice experiment. *Soc Sci Med*. 2019;233(May):28-37. doi:10.1016/j.socscimed.2019.05.032

106. Ratcliffe J, Cameron I, Lancsar E, et al. Developing a new quality of life instrument with older people for economic evaluation in aged care: study protocol. *BMJ Open*. 2019;9(5):e028647. doi:10.1136/bmjopen-2018-028647

107. Lorenzoni L, Marino A, Morgan D, James C. Health Spending Projections to 2030: New results based on a revised OECD methodology. *OECD Heal Work Pap*. 2019;(110). doi:10.1787/5667f23d-en

108. Papanicolas I, Marino A, Lorenzoni L, Jha A. Comparison of Health Care Spending by Age in 8 High-Income Countries. *JAMA Netw open*. 2020;3(8):e2014688. doi:10.1001/jamanetworkopen.2020.14688

109. Bulamu NB, Kaambwa B, Ratcliffe J. A systematic review of instruments for measuring outcomes in economic evaluation within aged care. *Health Qual Life Outcomes*. 2015;13(1):1-23. doi:10.1186/s12955-015-0372-8

110. Cleland J, Hutchinson C, Khadka J, Milte R, Ratcliffe J. A Review of the Development and Application of Generic Preference-Based Instruments with the Older Population. *Appl Health Econ Health Policy*. 2019;(0123456789). doi:10.1007/s40258-019-00512-4

111. Helter TM, Coast J, Łaszewska A, Stamm T, Simon J. *Capability Instruments in Economic Evaluations of Health-Related Interventions: A Comparative Review of the Literature*. Springer International Publishing; 2019. doi:10.1007/s11136-019-02393-5

112. Bowling A, Stenner P. Which measure of quality of life performs best in older age? A comparison of the OPQOL, CASP-19 and WHOQOL-OLD. *J Epidemiol Community Health*. 2011;65(3):273-280. doi:10.1136/jech.2009.087668

113. Hackert MQN, Exel J van, Brouwer WBF. Valid Outcome Measures in Care for Older People: Comparing the ASCOT and the ICECAP-O. *Value Heal*. 2017;20(7):936-944. doi:10.1016/j.jval.2017.03.012

114. Davis JC, Liu-Ambrose T, Richardson CG, Bryan S. A comparison of the ICECAP-O with EQ-5D in a falls prevention clinical setting: Are they complements or substitutes? *Qual Life Res*. 2013;22(5):969-977. doi:10.1007/s11136-012-0225-4

115. Van Leeuwen KM, Bosmans JE, Jansen APD, et al. Comparing measurement properties of the EQ-5D-3L, ICECAP-O, and ASCOT in frail older adults. *Value Heal*. 2015;18(1):35-43. doi:10.1016/j.jval.2014.09.006

116. Hackert MQN, van Exel J, Brouwer WBF. Content validation of the Well-being of Older People measure (WOOP). *Health Qual Life Outcomes*. 2021;19(1):200. doi:10.1186/s12955-021-01834-5

117. Hackert MQN, Brouwer WBF, Hoefman RJ, van Exel J. Views of older people in the Netherlands on wellbeing: A Q-methodology study. *Soc Sci Med*. 2019;240(December 2018):112535. doi:10.1016/j.socscimed.2019.112535

118. Bahrampour M, Byrnes J, Norman R, Scuffham PA, Downes M. Discrete choice experiments to generate utility values for multi - attribute utility instruments : a systematic review of methods. *Eur J Heal Econ*. Published online 2020. doi:10.1007/s10198-020-01189-6

119. Mulhern B, Bansback N, Brazier J, et al. Preparatory study for the revaluation of the EQ-5D tariff: methodology report. *Health Technol Assess (Rockv)*. 2014;18(12). doi:10.3310/hta18120

120. Dolan P, Stalmeier P. The validity of time trade-off values in calculating QALYs: constant proportional time trade-off versus the proportional heuristic. *J Health Econ*. 2003;22(3):445-458. doi:10.1016/S0167-6296(02)00120-0

121. Craig BM, Rand K, Bailey H, Stalmeier PFM. Quality-Adjusted Life-Years without Constant Proportionality. *Value Heal*. 2018;21(9):1124-1131. doi:10.1016/j.jval.2018.02.004

122. Jonker MF, Donkers B, de Bekker-Grob EW, Stolk EA. Advocating a Paradigm Shift in Health-State Valuations: The Estimation of Time-Preference Corrected QALY Tariffs. *Value Heal*. 2018;21(8):993-1001. doi:10.1016/j.jval.2018.01.016

123. Jonker MF, Norman R. Not all respondents use a multiplicative utility function in choice experiments for health state valuations, which should be reflected in the elicitation format (or statistical analysis). *Heal Econ (United Kingdom)*. 2022;31(2):431-439. doi:10.1002/hec.4457

124. Jonker MF, Bliemer MCJ. On the Optimization of Bayesian D-Efficient Discrete Choice Experiment Designs for the Estimation of QALY Tariffs That Are Corrected for Nonlinear Time Preferences. *Value*

Heal. 2019;22(10):1162-1169. doi:10.1016/j.jval.2019.05.014

125. Flynn TN, Bilger M, Malhotra C, Finkelstein EA. Are Efficient Designs Used in Discrete Choice Experiments Too Difficult for Some Respondents? A Case Study Eliciting Preferences for End-of-Life Care. *Pharmacoeconomics*. 2016;34(3):273-284. doi:10.1007/s40273-015-0338-z
126. Mulhern B, Bansback N, Hole AR, Tsuchiya A. Using Discrete Choice Experiments with Duration to Model EQ-5D-5L Health State Preferences: Testing Experimental Design Strategies. *Med Decis Mak*. 2017;37(3):285-297. doi:10.1177/0272989X16670616
127. Himmler S, Soekhai V, van Exel J, Brouwer W. What works better for preference elicitation among older people? Cognitive burden of discrete choice experiment and case 2 best-worst scaling in an online setting. *J Choice Model*. 2020;38(November 2020):100265. doi:10.1016/j.jocm.2020.100265
128. van Leeuwen KM, van Loon MS, van Nes FA, et al. What does quality of life mean to older adults? A thematic synthesis. Ginsberg SD, ed. *PLoS One*. 2019;14(3):e0213263. doi:10.1371/journal.pone.0213263
129. Douma L, Steverink N, Hutter I, Meijering L, Bowers BJ. Exploring subjective well-being in older age by using participant-generated word clouds. *Gerontologist*. 2017;57(2):229-239. doi:10.1093/geront/gnv119
130. Attema AE, Brouwer WBF, Claxton K. Discounting in Economic Evaluations. *Pharmacoeconomics*. 2018;36(7):745-758. doi:10.1007/s40273-018-0672-z
131. Mangen MJJ, Bolkenbaas M, Huijts SM, van Werkhoven CH, Bonten MJM, de Wit GA. Quality of life in community-dwelling Dutch elderly measured by EQ-5D-3L. *Health Qual Life Outcomes*. 2017;15(1):1-6. doi:10.1186/s12955-016-0577-5
132. Nijsten JMH, Leontjevas R, Smalbrugge M, Koopmans RTCM, Gerritsen DL. Apathy and health-related quality of life in nursing home residents. *Qual Life Res*. 2019;28(3):751-759. doi:10.1007/s11136-018-2041-y
133. Himmler S, van Exel J, Brouwer W. Estimating the monetary value of health and capability well-being applying the well-being valuation approach. *Eur J Heal Econ*. 2020;(0123456789). doi:10.1007/s10198-020-01231-7
134. Kinghorn P, Afentou N. Eliciting a monetary threshold for a year of sufficient capability to inform resource allocation decisions in public health and social care. *Soc Sci Med*. 2021;279(January):113977. doi:10.1016/j.socscimed.2021.113977
135. Nuzzo JB, Shearer MP. International Engagement Is Critical to Fighting Epidemics. *Heal Secur*. 2017;15(1):33-35. doi:10.1089/hs.2016.0098
136. World Bank Group. 2014-2015 West Africa Ebola Crisis: Impact Update. *World Bank Fisc Rep*. Published online 2016:4. http://pubdocs.worldbank.org/en/297531463677588074/Ebola-Economic-Impact-and-Lessons-Paper-short-version.pdf
137. Plass D, Mangen M-JJ, Kraemer A, et al. The disease burden of hepatitis B, influenza, measles and salmonellosis in Germany: first results of the Burden of Communicable Diseases in Europe Study. *Epidemiol Infect*. 2014;142(10):2024-2035. doi:10.1017/S0950268813003312
138. Cassini A, Colzani E, Pini A, et al. Impact of infectious diseases on population health using incidence-based disability-adjusted life years (DALYs): results from the Burden of Communicable Diseases in Europe study , European Union and European Economic Area countries , 2009 to 2013. *Euro Surveill*. 2018;23(16).
139. Perry-Duxbury M, van Exel J, Brouwer W. How to value safety in economic evaluations in health care? A review of applications in different sectors. *Eur J Heal Econ*. 2019;20(7):1041-1061. doi:10.1007/s10198-019-01076-9
140. Alberini A, Hunt A, Markandya A. Willingness to pay to reduce mortality risks: Evidence from a three-country contingent valuation study. *Environ Resour Econ*. 2006;33(2):251-264. doi:10.1007/s10640-005-3106-2
141. Corso P, Ingels J, Roldos M. A Comparison of Willingness to Pay to Prevent Child Maltreatment Deaths in Ecuador and the United States. *Int J Environ Res Public Health*. 2013;10(4):1342-1355. doi:10.3390/ijerph10041342
142. Dealy BC, Horn BP, Callahan TJ, Bryan AD. The economic impact of Project MARS (Motivating Adolescents to Reduce Sexual Risk). *Heal Psychol*. 2013;32(9):1003-1012. doi:10.1037/a0033607
143. Determann D, Korfage IJ, Lambooij MS, et al. Acceptance of Vaccinations in Pandemic Outbreaks: A Discrete Choice Experiment. Reddy J, ed. *PLoS One*. 2014;9(7):e102505. doi:10.1371/journal.pone.0102505
144. Perry-Duxbury M, van Exel J, Brouwer W. The value of safety: A literature review. *Forthcoming*.
145. Hofstede G. National Cultures in Four Dimensions: A Research-Based Theory of Cultural Differences among Nations. *Int Stud Manag Organ*. 1983;13(1-2):46-74. doi:10.1080/00208825.1983.11656358
146. Eurostat. GDP per capita in PPS. Index (EU28 = 100). Published 2018. Accessed March 15, 2018. https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tec00114&plugi

n=1

147. Bobinac A, Van Exel NJA, Rutten FFH, Brouwer WBF. Willingness to pay for a quality-adjusted life-year: The individual perspective. *Value Heal*. 2010;13(8):1046-1055. doi:10.1111/j.1524-4733.2010.00781.x

148. Johnson FR, Banzhaf MR, Desvousges WH. Willingness to pay for improved respiratory and cardiovascular health: a multiple-format, stated-preference approach. *Health Econ*. 2000;9(4):295-317. http://www.ncbi.nlm.nih.gov/pubmed/10862074

149. Donaldson C, Thomas R, Torgerson DJ. Validity of open-ended and payment scale approaches to eliciting willingness to pay. *Appl Econ*. 1997;29(1):79-84. doi:10.1080/000368497327425

150. Blomquist GC, Blumenschein K, Johannesson M. Eliciting willingness to pay without bias using follow-up certainty statements: Comparisons between probably/definitely and a 10-point certainty Scale. *Environ Resour Econ*. 2009;43(4):473-502. doi:10.1007/s10640-008-9242-8

151. Huls SPI, van Osch SMC, Brouwer WBF, van Exel J, Stiggelbout AM. Psychometric evaluation of the Health-Risk Attitude Scale (HRAS-13): assessing the reliability, dimensionality and validity in the general population and a patient population. *Psychol Heal*. 2020;0(0):1-17. doi:10.1080/08870446.2020.1851689

152. Marozzi M. Measuring Trust in European Public Institutions. *Soc Indic Res*. 2015;123(3):879-895. doi:10.1007/s11205-014-0765-9

153. Simonson I, Drolet A. Anchoring Effects on Consumers' Willingness-to-Pay and Willingness-to-Accept. *J Consum Res*. 2004;31(3):681-690. doi:10.1086/425103

154. Gyrd-Hansen D, Jensen ML, Kjaer T. Framing the willingness-to-pay question : Impact on response pattern and mean willingness to pay. *Health Econ*. 2014;23(5):550-563. doi:10.1002/hec.2932

155. Frew EJ, Whynes DK, Wolstenholme JL. Eliciting Willingness to Pay: Comparing Closed-Ended with Open-Ended and Payment Scale Formats. *Med Decis Mak*. 2003;23(2):150-159. doi:10.1177/0272989X03251245

156. de Bekker-Grob EW, Polder JJ, Mackenbach JP, Meerding WJ. Towards a comprehensive estimate of national spending on prevention. *BMC Public Health*. 2007;7(1):252. doi:10.1186/1471-2458-7-252

157. Bloom D, Kuhn M, Prettner K. Modern Infectious Diseases: Macroeconomic Impacts and Policy Responses. *NBER Work Pap*. 2020;No. 27757.

158. Morens DM, Fauci AS. Emerging Pandemic Diseases: How We Got To COVID-19. *Cell*. 2020;182(5):1077-1092. doi:10.1016/j.cell.2020.08.021

159. Dobson AP, Pimm SL, Hannah L, et al. Ecology and economics for pandemic prevention (supplementary material). *Science (80-)*. 2020;369(6502):379-381. doi:10.1126/science.abc3189

160. Morse SS, Mazet JAK, Woolhouse M, et al. Prediction and prevention of the next pandemic zoonosis. *Lancet*. 2012;380(9857):1956-1965. doi:10.1016/S0140-6736(12)61684-5

161. Osterhaus A, Mackenzie J. Pandemic preparedness planning in peacetime: what is missing? *One Heal Outlook*. 2020;2(1):20-23. doi:10.1186/s42522-020-00027-2

162. Chilton S, Nielsen JS, Wildman J. Beyond COVID-19: How the 'dismal science' can prepare us for the future. *Heal Econ (United Kingdom)*. 2020;(May):1-3. doi:10.1002/hec.4114

163. Himmler S, van Exel J, Perry-Duxbury M, Brouwer W. Willingness to pay for an early warning system for infectious diseases. *Eur J Heal Econ*. Published online 2020. doi:10.1007/s10198-020-01171-2

164. Kling CL, Phaneuf DJ, Zhao J. From Exxon to BP: Has Some Number Become Better than No Number? *J Econ Perspect*. 2012;26(4):3-26. doi:10.1257/jep.26.4.3

165. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: An EQ-5D-5L value set for England. *Heal Econ (United Kingdom)*. 2018;27(1):7-22. doi:10.1002/hec.3564

166. Bobinac A, van Exel J, Rutten FFH, Brouwer WBF. The Value of a QALY: Individual Willingness to Pay for Health Gains Under Risk. *Pharmacoeconomics*. 2014;32(1):75-86. doi:10.1007/s40273-013-0110-1

167. HIQA. Review of restrictive public policy measures to limit the spread of COVID-19. Published 2020. Accessed June 5, 2020. https://www.hiqa.ie/reports-and-publications/health-technology-assessment/review-restrictive-public-policy-measures

168. ECDC. *Data on the Geographic Distribution of COVID-19 Cases Worldwide*.; 2020. Accessed June 5, 2020. https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

169. Hale T, Webster S, Petherick A, Phillips T, Beatrz K. *Oxford COVID-19 Government Response Tracker*. Blavatnik School of Government. Data use policy: Creative Commons Attribution CC BY standard.; 2020.

170. Eurostat. Harmonised index of consumer prices - monthly data. Published 2020. Accessed June 9, 2020. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=prc_hicp_midx&lang=en

171. Huls S, van Osch S, Brouwer W, van Exel J, Stiggelbout A. Psychometric evaluation of the Health-Risk Attitude Scale (HRAS-13): Assessing the reliability, dimensionality and validity in the general population and a patient population. *Psychol Heal*. Published online 2020. doi:10.1080/08870446.2020.1851689

172. Weesie J. Seemlingly unrelated estimation and the cluster-adjusted sandwich estimator. *Stata Tech Bull*. 2000;(52):34-47.

173. Kritikos VAS, Graeber D. Corona-Pandemie wird zur Krise für Selbständige. *DIW aktuell*. 2020;2020(47).

174. European Parliament. *Public Opinion Monitoring at a Glance in the Time of COVID-19*.; 2020. https://www.europarl.europa.eu/at-your-service/files/be-heard/eurobarometer/2020/covid19/en-public-opinion-in-the-time-of-COVID19-27052020.pdf

175. Sabat I, Neuman-Böhme S, Varghese NE, et al. United but divided: Policy responses and people's perceptions in the EU during the COVID-19 outbreak. *Health Policy (New York)*. Published online 2020. doi:10.1016/j.healthpol.2020.06.009

176. Neumann-Böhme S, Varghese NE, Sabat I, et al. Once we have it, will we use it? A European survey on willingness to be vaccinated against COVID-19. *Eur J Heal Econ*. 2020;21(7):977-982. doi:10.1007/s10198-020-01208-6

177. Entringer T, Kröger H. Einsam , aber resilient – Die Menschen haben den Lockdown besser verkraftet als vermutet. *DIW aktuell*. 2020;(46). https://www.diw.de/documents/publikationen/73/diw_01.c.791373.de/diw_aktuell_46.pdf

178. List JA, Gallet CA. What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values? *Environ Resour Econ*. 2001;20(3):241-254. doi:10.1023/A:1012791822804

179. Yu F, Geldsetzer P, Meierkord A, et al. Knowledge About COVID-19 Among Adults in China: Cross-sectional Online Survey. *J Med Internet Res*. 2021;23(4):e26940. doi:10.2196/26940

180. Geldsetzer P. Use of Rapid Online Surveys to Assess People's Perceptions During Infectious Disease Outbreaks: A Cross-sectional Survey on COVID-19. *J Med Internet Res*. 2020;22(4):e18790. doi:10.2196/18790

181. Chen S, Chen Q, Yang J, et al. Curbing the COVID-19 pandemic with facility-based isolation of mild cases: a mathematical modeling study. *J Travel Med*. 2021;28(2). doi:10.1093/jtm/taaa226

182. Chen S, Zhang Z, Yang J, et al. Fangcang shelter hospitals: a novel concept for responding to public health emergencies. *Lancet*. 2020;395(10232):1305-1314. doi:10.1016/S0140-6736(20)30744-3

183. Chia ML, Him Chau DH, Lim KS, Yang Liu CW, Tan HK, Tan YR. Managing COVID-19 in a Novel, Rapidly Deployable Community Isolation Quarantine Facility. *Ann Intern Med*. 2021;174(2):247-251. doi:10.7326/M20-4746

184. Sen A. Capability and Well-being. In: *The Quality of Life*. In M. C. Nussbaum (Ed.), The quality of life. Oxford: Clarendon Press.; 1993.

185. Engel L, Mortimer D, Bryan S, Lear SA, Whitehurst DGT. An Investigation of the Overlap Between the ICECAP-A and Five Preference-Based Health-Related Quality of Life Instruments. *Pharmacoeconomics*. 2017;35(7):741-753. doi:10.1007/s40273-017-0491-7

186. Mitchell PM, Al-Janabi H, Byford S, et al. Assessing the validity of the ICECAP-A capability measure for adults with depression. *BMC Psychiatry*. 2017;17(1):1-13. doi:10.1186/s12888-017-1211-8

187. Mitchell PM, Al-Janabi H, Richardson J, Iezzi A, Coast J. The relative impacts of disease on health status and capability wellbeing: A multi-country study. *PLoS One*. 2015;10(12):1-15. doi:10.1371/journal.pone.0143590

188. Chen G, Ratcliffe J, Kaambwa B, McCaffrey N, Richardson J. Empirical Comparison Between Capability and Two Health-Related Quality of Life Measures. *Soc Indic Res*. Published online 2017. doi:10.1007/s11205-017-1788-9

189. Keeley T, Coast J, Nicholls E, Foster NE, Jowett S, Al-Janabi H. An analysis of the complementarity of ICECAP-A and EQ-5D-3 L in an adult population of patients with knee pain. *Health Qual Life Outcomes*. 2016;14(1):1-5. doi:10.1186/s12955-016-0430-x

190. Versteegh M, Knies S, Brouwer W. From Good to Better: New Dutch Guidelines for Economic Evaluations in Healthcare. *Pharmacoeconomics*. 2016;34(11):1071-1074. doi:10.1007/s40273-016-0431-y

191. Karimi M, Brazier J, Basarir H. The Capability Approach: A Critical Review of Its Application in Health Economics. *Value Heal*. 2016;19(6):795-799. doi:10.1016/j.jval.2016.05.006

192. Dolan P, Fujiwara D. Happiness-Based Policy Analysis. In: Adler MD, Fleurbaey M, eds. *The Oxford Handbook of Well-Being and Public Policy*. Vol 1. Oxford University Press; 2016:1-41. doi:10.1093/oxfordhb/9780199325818.013.9

193. Dolan P. Developing methods that really do value the "Q" in the QALY. *Heal Econ Policy Law*. 2008;3(1):69-77. doi:10.1017/S1744133107004355

194. Hausman J. Contingent Valuation: From Dubious to Hopeless. *J Econ Perspect*. 2012;26(4):43-56. doi:10.1257/jep.26.4.43

195. Veenhoven R. Capability and happiness: Conceptual difference and reality links. *J Socio Econ*. 2010;39(3):344-350. doi:10.1016/j.socec.2009.11.007

196. Engel L, Bryan S, Noonan VK, Whitehurst DGT. Using path analysis to investigate the relationships

between standardized instruments that measure health-related quality of life, capability wellbeing and subjective wellbeing: An application in the context of spinal cord injury. *Soc Sci Med*. 2018;213(January):154-164. doi:10.1016/j.socscimed.2018.07.041

197. Fujiwara D. A General Method for Valuing Non-Market Goods Using Wellbeing Data : Three-Stage Wellbeing Valuation. *Cent Econ Perform Discuss Pap No 1233*. 2013;(1233). doi:http://cep.lse.ac.uk/pubs/download/dp1233.pdf
198. Huls S, van Osch S, Brouwer W, van Exel NJA, Stiggelbout AM. Development and validation of the Health-Risk Attitude Scale. *Forthcoming*. Published online 2018.
199. Garrido S, Méndez I, Abellán JM. Analysing the Simultaneous Relationship Between Life Satisfaction and Health-Related Quality of Life. *J Happiness Stud*. 2013;14(6):1813-1838. doi:10.1007/s10902-012-9411-x
200. Brown TT. The Subjective Well-Being Method of Valuation: An Application to General Health Status. *Health Serv Res*. 2015;50(6):1996-2018. doi:10.1111/1475-6773.12294
201. Dolan P, Peasgood T, White M. Do we really know what makes us happy? A review of the economic literature on the factors associated with subjective well-being. *J Econ Psychol*. 2008;29(1):94-122. doi:10.1016/J.JOEP.2007.09.001
202. Fujiwara D, Campbell R. *Valuation Techniques for Social Cost-Benefit Analysis: Stated Preference, Revealed Preference and Subjective Well-Being Approaches*.; 2011.
203. Steel P, Schmidt J, Shultz J. Refining the relationship between personality and subjective well-being. *Psychol Bull*. 2008;134(1):138-161. doi:10.1037/0033-2909.134.1.138
204. Schyns P. Income and Satisfaction in Russia. *J Happiness Stud*. 2001;2(2):173-204. doi:10.1023/A:1011564631319
205. Diener E, Lucas RE, Oishi S, Suh EM. Looking Up and Looking Down: Weighting Good and Bad Information in Life Satisfaction Judgments. *Personal Soc Psychol Bull*. 2002;28(4):437-445. doi:10.1177/0146167202287002
206. Howley P. Valuing the benefits from health care interventions using life satisfaction data. *Heal Econom Data Gr Work Pap*. 2016;1(January).
207. Luttmer E. Neighbors as Negatives: Relative Earnings and Well-Being. *Q J Econ*. 2005;(August):51. doi:10.3386/w10667
208. Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data*. MIT Press; 2010.
209. Layard R, Nickell S, Mayraz G. The marginal utility of income. *J Public Econ*. 2008;92(8-9):1846-1857. doi:10.1016/j.jpubeco.2008.01.007
210. Norman GR, Sloan JA, Wyrwich KW. Interpretation of Changes in Health-related Quality of Life: the remarkable universality of half a standard deviation. *Med Care*. 2003;41(5):582-592. doi:10.1097/01.MLR.0000062554.74615.4C
211. Baum C, Schaffer M, Stillman S. ivreg2: Stata module for extended instrumental variables/2SLS, GMM and AC/HAC, LIML and k-class regression. Published 2010. http://ideas.repec.org/c/boc/bocode/s425401.html
212. Van Hout B, Janssen MF, Feng YS, et al. Interim scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Heal*. 2012;15(5):708-715. doi:10.1016/j.jval.2012.02.008
213. Al-Janabi H, Flynn TN, Peters TJ, Bryan S, Coast J. Test-Retest Reliability of Capability Measurement in the UK General Population. *Health Econ*. 2015;24(5):625-630. doi:10.1002/hec.3100
214. Baker R, Bateman I, Donaldson C, et al. Weighting and valuing quality-adjusted life-years using stated preference methods: Preliminary results from the social value of a QALY project. *Health Technol Assess (Rockv)*. 2010;14(27). doi:10.3310/hta14270
215. Diener E, Oishi S, Tay L. Advances in subjective well-being research. *Nat Hum Behav*. 2018;2(4):253-260. doi:10.1038/s41562-018-0307-6
216. Veenhoven R. Happiness, Also Known as "Life Satisfaction" and "Subjective Well-Being." In: Land KC, Michalos AC, Sirgy MJ, eds. *Handbook of Social Indicators and Quality of Life Research*. Springer Netherlands; 2012. doi:10.1007/978-94-007-2421-1
217. Hofman C, Makai P, Boter H, et al. The influence of age on health valuations: the older olds prefer functional independence while the younger olds prefer less morbidity. *Clin Interv Aging*. 2015;10:1131. doi:10.2147/CIA.S78698
218. Claxton K, Martin S, Soares M, et al. Methods for the estimation of the National Institute for Health and care excellence cost-effectiveness threshold. *Health Technol Assess (Rockv)*. 2015;19(14):1-503. doi:10.3310/hta19140
219. Kinghorn P. Using deliberative methods to establish a sufficient state of capability well-being for use in decision-making in the contexts of public health and social care. *Soc Sci Med*. 2019;240(August 2018):112546. doi:10.1016/j.socscimed.2019.112546
220. Donaldson C, Mitton C. Coronavirus: Where has all the health economics gone? *Int J Heal Policy Manag*. 2020;9(11):466-468. doi:10.34172/ijhpm.2020.108

221. Hendren N, Sprung-Keyser B. A unified welfare analysis of government policies. *Q J Econ*. 2020;135(3):1209-1318.

222. McIntosh E. Introduction. In: McIntosh E, Clarke P, Frew E, Louviere J, eds. *Applied Methods of Cost–Benefit Analysis in Health Care*. Oxford University Press; 2010:3-18.

223. Rowen D, Azzabi Zouraq I, Chevrou-Severac H, van Hout B. International Regulations and Recommendations for Utility Data for Health Technology Assessment. *Pharmacoeconomics*. 2017;35(s1):11-19. doi:10.1007/s40273-017-0544-y

224. Neumann PJ, Sanders GD, Russell LB, Siegel JE, Ganiats TG. *Cost Effectiveness in Health and Medicine*. Oxford University Press; 2016.

225. Cleemput I, Neyt M, Thiry N, De Laet C, Leys M. Using threshold values for cost per quality-adjusted life-year gained in healthcare decisions. *Int J Technol Assess Health Care*. 2011;27(1):71-76. doi:10.1017/S0266462310001194

226. Bundesministerium für Gesundheit. Stellungnahme zur Methodik der Kosten-Nutzen-Bewertung von Arzneimitteln. Published 2008. http://www.bmg.bund.de/

227. Ferrer-i-Carbonell A, van Praag BMS. The subjective costs of health losses due to chronic diseases. An alternative model for monetary appraisal. *Health Econ*. 2002;11(8):709-722. doi:10.1002/hec.696

228. Howley P. Less money or better health? Evaluating individual's willingness to make trade-offs using life satisfaction data. *J Econ Behav Organ*. 2017;135:53-65. doi:10.1016/j.jebo.2017.01.010

229. McNamee P, Mendolia S. Changes in health-related quality of life: a compensating income variation approach. *Appl Econ*. 2018;00(00):1-12. doi:10.1080/00036846.2018.1504160

230. Mcdonald R, Powdthavee N. The Shadow Prices of Voluntary Caregiving : Using Panel Data of Well-Being to Estimate the Cost of Informal Care. *IZA Discuss Pap Ser*. 2018;(11545).

231. van den Berg B, Ferrer-i-Carbonell A. Monetary valuation of informal care: the well-being valuation method. *Health Econ*. 2007;16(11):1227-1244. doi:10.1002/hec.1224

232. Luechinger S. Valuing Air Quality Using the Life Satisfaction Approach. *Econ J*. 2009;119:482-515. doi:10.1111/j.1468-0297.2008.02241.x

233. Luechinger S, Raschky PA. Valuing flood disasters using the life satisfaction approach. *J Public Econ*. 2009;93(3-4):620-633. doi:10.1016/j.jpubeco.2008.10.003

234. Frey BS, Luechinger S, Stutzer A. The life satisfaction approach to valuing public goods: The case of terrorism. *Public Choice*. 2009;138(3-4):317-345. doi:10.1007/s11127-008-9361-3

235. Dolan P, Kavetsos G, Krekel C, et al. Quantifying the intangible impact of the Olympics using subjective well-being data. *J Public Econ*. 2019;177:104043. doi:10.1016/j.jpubeco.2019.07.002

236. Himmler S, van Exel J, Brouwer W. Estimating the monetary value of health and capability well-being applying the well-being valuation approach. *Eur J Heal Econ*. 2020;21:1235-1244. doi:10.1007/s10198-020-01231-7

237. Bobinac A, van Exel NJA, Rutten FFH, Brouwer WBF. Valuing QALY gains by applying a societal perspective. *Health Econ*. 2013;22(10):1272-1281. doi:10.1002/hec.2879

238. Dolan P, Olsen JA. Equity in health: the importance of different health streams. *J Health Econ*. 2001;20(5):823-834. doi:10.1016/S0167-6296(01)00095-9

239. Pinto-Prades J-L, Sánchez-Martínez F-I, Corbacho B, Baker R. Valuing QALYs at the end of life. *Soc Sci Med*. 2014;113:5-14. doi:10.1016/j.socscimed.2014.04.039

240. Wagstaff A. QALYs and the equity-efficiency trade-off. *J Health Econ*. 1991;10(1):21-41. doi:10.1016/0167-6296(91)90015-F

241. Gyrd-Hansen D. Willingness to pay for a QALY. *Health Econ*. 2003;12(12):1049-1060. doi:10.1002/hec.799

242. Philipson TJ, Jena AB. Who Benefits From New Medical Technologies ? Estimates of consumer and producer surpluses for HIV/AIDS drugs. *NBER Work Pap*. 2005;11810.

243. Soeteman L, van Exel J, Bobinac A. The impact of the design of payment scales on the willingness to pay for health gains. *Eur J Heal Econ*. 2017;18(6):743-760. doi:10.1007/s10198-016-0825-y

244. Bayer C, Juessen F. Happiness and the persistence of income shocks. *Am Econ J Macroecon*. 2015;7(4):160-187. doi:10.1257/mac.20120163

245. Cai S, Park A. Permanent income and subjective well-being. *J Econ Behav Organ*. 2016;130:298-319. doi:10.1016/j.jebo.2016.07.016

246. Hariri JG, Lassen DD. Income and Outcomes: Social Desirability Bias Distorts Measurements of the Relationship between Income and Political Behavior. *Public Opin Q*. 2017;81(2):564-576. doi:10.1093/poq/nfw044

247. Rousseeuw P, Leroy A. *Robust Regression and Outlier Detection (Vol. 1)*. Wiley Online Library; 1987.

248. Verardi V, Croux C. Robust Regression in Stata. *Stata J Promot Commun Stat Stata*. 2009;9(3):439-453. doi:10.1177/1536867X0900900306

249. Bollen KA, Jackman RW. Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases. *Sociol Methods Res*. 1985;13(4):510-542. doi:10.1177/0049124185013004004

250. Ólafsdóttir T, Ásgeirsdóttir TL, Norton EC. Valuing pain using the subjective well-being method. *Econ Hum Biol*. 2020;37. doi:10.1016/j.ehb.2019.100827

251. Finkelstein A, Luttmer EFP, Notowidigdo MJ. What good is health without wealth? The effect of health on the marginal utility of consumption. *J Eur Econ Assoc*. 2013;11:221-258. doi:10.1111/j.1542-4774.2012.01101.x

252. Kools L, Knoef M. Health and consumption preferences; estimating the health state dependence of utility using equivalence scales. *Eur Econ Rev*. 2019;113:46-62. doi:10.1016/j.euroecorev.2018.12.007

253. Ware J, Keller S, Kosinski M. *Sf-12: How to Score the Sf-12 Physical and Mental Health Summary Scales*. Health Institute, New England Medical Center; 1995.

254. Goebel J, Grabka M, Liebig S, et al. The german socio-economic panel (soep). *Jahrb Natl Okon Stat*. 2019;239(2):345-360.

255. Hagenaars A, De Vos K, M AZ. *Poverty Statistics in the Late 1980s: Research Based on Micro-Data.* Office for Official Publications of the European Communities.; 1994.

256. Federal Statistical Office. Consumer price index (incl. rates of change): Germany, years. Table: 61111-0001.

257. Pischke J-S. MONEY AND HAPPINESS: EVIDENCE FROM THE INDUSTRY WAGE STRUCTURE. *NBER Work Pap*. 2011;(17056).

258. Frijters P, Haisken-Denew JP, Shields MA. Money Does Matter! Evidence from Increasing Real Income and Life Satisfaction in East Germany Following Reunification. *Am Econ Rev*. 2004;94(3):730-740. doi:10.1257/0002828041464551

259. Vatter J. Well-being in Germany: What explains the regional variation? *Discuss Pap Forschungszentrum Gener der Albert-Ludwigs- Univ Freibg*. 2012;No. 50.

260. Norman R, Cronin P, Viney R, King M, Street D, Ratcliffe J. International Comparisons in Valuing EQ-5D Health States : *Value Heal*. 2009;12(8):1194-1200. doi:10.1111/j.1524-4733.2009.00581.x

261. Woods B, Revill P, Sculpher M, Claxton K. Country-Level Cost-Effectiveness Thresholds: Initial Estimates and the Need for Further Research. *Value Heal*. 2016;19(8):929-935. doi:10.1016/j.jval.2016.02.017

262. Ochalek J, Lomas J. Reflecting the Health Opportunity Costs of Funding Decisions Within Value Frameworks: Initial Estimates and the Need for Further Research. *Clin Ther*. 2020;42(1):44-59.e2. doi:10.1016/j.clinthera.2019.12.002

263. Pischke J-S. A CAUTIONARY NOTE ON USING INDUSTRY AFFILIATION TO PREDICT INCOME. *NBER Work Pap Ser A*. 2012;(18384).

264. Borghans L, Duckworth AL, Heckman JJ, Weel B ter. The Economics and Psychology of Personality Traits. *J Hum Resour*. 2008;43(4):972-1059. doi:10.3368/jhr.43.4.972

265. Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. *J Am Stat Assoc*. 1996;91(434):444. doi:10.2307/2291629

266. Ambrosio CD, Clark A, Zhu R. Living in the Shadow of the Past : Financial Profiles , Health and Well-Being Living in the Shadow of the Past : Financial Profiles , Health and Well-Being ∗. *Work Pap*. Published online 2018. http://www.iariw.org/copenhagen/dambrosio.pdf

267. Boyce CJ, Wood AM, Banks J, Clark AE, Brown GDA. Money, Well-Being, and Loss Aversion. *Psychol Sci*. 2013;24(12):2557-2562. doi:10.1177/0956797613496436

268. Sabatini F. The relationship between happiness and health: Evidence from Italy. *Soc Sci Med*. 2014;114:178-187. doi:10.1016/j.socscimed.2014.05.024

269. Loewenstein G, Ubel PA. Hedonic adaptation and the role of decision and experience utility in public policy. *J Public Econ*. 2008;92(8-9):1795-1810. doi:10.1016/j.jpubeco.2007.12.011

270. Oswald AJ, Powdthavee N. Does happiness adapt? A longitudinal study of disability with implications for economists and judges. *J Public Econ*. 2008;92(5-6):1061-1077. doi:10.1016/J.JPUBECO.2008.01.002

271. Etilé F, Frijters P, Johnston D, Shields M. *Psychological Resilience to Major Socioeconomic Life Events*.; 2020.

272. Mogstad M, Santos A, Torgovitsky A. Using Instrumental Variables for Inference About Policy Relevant Treatment Parameters. *Econometrica*. 2018;86(5):1589-1619. doi:10.3982/ECTA15463

273. Cookson R, Skarda I, Cotton-Barratt O, Adler M, Asaria M, Ord T. Quality adjusted life years based on health and consumption: A summary wellbeing measure for cross-sectoral economic evaluation. *Heal Econ (United Kingdom)*. 2021;30(1):70-85. doi:10.1002/hec.4177

274. Herrera-Araujo D, Hammitt JK, Rheinberger CM. Theoretical bounds on the value of improved health. *J Health Econ*. 2020;72:102341. doi:10.1016/j.jhealeco.2020.102341

275. Ryan M, Mentzakis E, Matheson C, Bond C. Survey modes comparison in contingent valuation: Internet panels and mail surveys. *Heal Econ (United Kingdom)*. 2019;(September 2019):234-242. doi:10.1002/hec.3983

276. Cornesse C, Bosnjak M. Is there an association between survey characteristics and representativeness? A meta-analysis. *Surv Res Methods*. 2018;12(1):1-13.

doi:10.18148/srm/2018.v12i1.7205

277. Huls SPI, Whichello CL, van Exel J, Uyl-de Groot CA, de Bekker-Grob EW. What Is Next for Patient Preferences in Health Technology Assessment? A Systematic Review of the Challenges. *Value Heal*. 2019;22(11):1318-1328. doi:10.1016/j.jval.2019.04.1930

278. Verma N, Shiroma K, Rich K, Fleischmann KR, Xie B, Lee MK. Conducting Quantitative Research with Hard-To-Reach-Online Populations: Using Prime Panels to Rapidly Survey Older Adults During a Pandemic. In: *Diversity, Divergence, Dialogue*. ; 2021:384-393. doi:10.1007/978-3-030-71305-8_32

279. van Baal P, Perry-Duxbury M, Bakx P, Versteegh M, van Doorslaer E, Brouwer W. A cost-effectiveness threshold based on the marginal returns of cardiovascular hospital spending. *Heal Econ (United Kingdom)*. 2019;28(1):87-100. doi:10.1002/hec.3831

280. Hoefman RJ, van Exel J, Brouwer WBF. Measuring Care-Related Quality of Life of Caregivers for Use in Economic Evaluations: CarerQol Tariffs for Australia, Germany, Sweden, UK, and US. *Pharmacoeconomics*. 2017;35(4):469-478. doi:10.1007/s40273-016-0477-x

281. Kangas O, Jauhiainen S, Simanainen M, Ylikännö M. The basic income experiment 2017–2018 in Finland. Preliminary results. *Reports Memo Minist Soc Aff Heal*. 2019;(9):34. http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161361/Report_The Basic Income Experiment 20172018 in Finland.pdf%0Ahttp://urn.fi/URN:ISBN:978-952-00-4035-2

282. de Vries LM, van Baal PHM, Brouwer WBF. Future Costs in Cost-Effectiveness Analyses: Past, Present, Future. *Pharmacoeconomics*. 2019;37(2):119-130. doi:10.1007/s40273-018-0749-8

283. Cubi-Molla P, Buxton M, Devlin N. Allocating Public Spending Efficiently: Is There a Need for a Better Mechanism to Inform Decisions in the UK and Elsewhere? *Appl Health Econ Health Policy*. 2021;(0123456789). doi:10.1007/s40258-021-00648-2