SOEP The Socio-Economic Panel

1101 2020

SOEPpapers

on Multidisciplinary Panel Data Research

SOEP – The German Socio-Economic Panel at DIW Berlin 1101-2020

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

Sebastian Himmler, Jannis Stöckel, Job van Exel, Werner Brouwer

DIW SOEP

ISSN: 1864-6689 (online)

German Socio-Economic Panel (SOEP)
DIW Berlin
Mohrenstrasse 58
10117 Berlin, Germany

Contact: soeppapers@diw.de

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

Sebastian Himmler*†, Jannis Stöckel*†, Job van Exel†‡ and Werner Brouwer†‡

Wednesday 29th July, 2020

Abstract

Cost-utility analysis compares the monetary cost of health interventions to the associated health consequences expressed using quality-adjusted life years (QALYs). At which threshold the ratio of both is still acceptable is a highly contested issue. Obtaining societal valuations of the monetary value of a QALY can help in setting such threshold values but it remains methodologically challenging. A recent study applied the well-being valuation approach to calculate such a monetary value using a compensating income variation approach. We explore the feasibility of this approach in a different context, using large-scale panel data from Germany. We investigate several important empirical and conceptual challenges such as the appropriate functional specification of income and the health state dependence of consumption utility. The estimated monetary values range from €20,000-60,000 with certain specifications leading to considerable deviations, underlining persistent practical challenges when applying the well-being valuation methodology to QALYs. Recommendations for future applications are formulated.

Keywords: Quality-adjusted life years, health valuation, well-being valuation, panel data, instrumental variable regression, piecewise regression

JEL Classification: D61, I18, I31, C33, C36

Acknowledgements: We would like to thank participants and especially discussants at the Nordic Health Economists' study group meeting 2019, the EuHEA PhD conference 2019, the 2019 International Health Economics Congress and the 2019 meeting of the German Health Economics Association. We further would like to thank Dorte Gyrd-Hansen and Tom Stargardt for excellent comments on a previous version of this paper. Sebastian Himmler receives funding from a Marie Sklodowska-Curie fellowship financed by the European Commission (Grant agreement No. 721402) and Jannis Stöckel receives funding from the *Smarter Choices for Better Health Initiative* of the Erasmus University Rotterdam. All remaining errors are our own.

*Corresponding author: Sebastian Himmler (himmler@eshpm.eur.nl) & Jannis Stöckel (stockel@eshpm.eur.nl)

†Erasmus School of Health Policy & Management, Erasmus University Rotterdam, Netherlands.

‡Erasmus School of Economics, Erasmus University Rotterdam, Netherlands.

1 Introduction

Public health care budgets are under increased strain by costly new health technologies, adding to the pressure from an ageing populations' expanding care demand (de Meijer et al., 2013). To allocate the available resources efficiently health authorities have to identify criteria that guide their reimbursement decisions to (ideally) reflect a set of implicit and explicit societal preferences. Along with clinical or ethical criteria, assessing whether a novel intervention offers appropriate value for money is of crucial importance in this context. In many jurisdictions this assessment is typically operationalised using cost-utility analysis (Rowen et al., 2017), where the costs of a new technology are compared to the expected health gain it generates, often measured using Quality Adjusted Life Years (QALYs) (Neumann et al., 2016). Equation (1) formulates the corresponding (simplified) decision rule, with ΔQ denoting the health gain (in QALYs) and Δc_t the total costs compared to the alternative treatment:

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

Taking a societal perspective, this ratio, also called the incremental cost-effectiveness ratio (ICER), is acceptable if it lies below v_Q, the consumption value of a QALY (Brouwer et al., 2019). While the use and empirical foundation of such threshold values vary across jurisdictions (Cameron et al., 2018; Cleemput et al., 2011), estimating the appropriate level of v_Q is inherently difficult. One way to obtain v_Q relies on stated preferences by asking individuals directly about their willingness to pay (WTP) for specific health gains. Ryen and Svensson (2015) summarised the large existing 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) for a gain in one QALY.

In a recent study, Huang et al. (2018) proposed an alternative method for estimating v_Q, which does not rely on stated preferences but on revealed, although subjective, information: the well-being valuation approach. This method has been applied before to obtain monetary valuations for various other non-market goods including specific health outcomes and diseases (Brown, 2015; Ferrer-i Carbonell & van Praag, 2002; Howley, 2017; McNamee & Mendolia, 2018), the provision of informal care (Mcdonald & Powdthavee, 2018; van den Berg & Ferrer-i Carbonell, 2007), air pollution (Luechinger, 2009), utility losses from natural disasters (Luechinger & Raschky, 2009), national security (Frey et al., 2009) or the welfare effects of international sports events (Dolan et al., 2019). Huang et al. (2018) extended this list to the valuation of QALYs by using data from the HILDA panel survey from Australia, obtaining v_Q estimates of A$42,000 (€28,000) to A$67,000 (€45,000), which were similar to threshold values applied for funding decisions in Australia.

Both stated preference WTP and well-being valuation approaches have clear advantages and disadvantages and may answer different questions based on how v_Q is specified. Stated preference methods allow researchers to tailor their experimental design to specific contexts and thereby elicit exactly what they want to include in the valuation of a QALY, while controlling for undesired influences. This for example includes expressing WTP from an individual or a societal perspective (Bobinac et al., 2013), thereby capturing more than self-interested motivations when establishing WTP-based estimates for v_Q. Similarly, equity concerns relating to specific health states or streams (Dolan & Olsen, 2001; Pinto-Prades et al., 2014), but also socio-economic health inequalities (Wagstaff, 1991) can be connected with the QALY framework. 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 (Gyrd-Hansen, 2003; Philipson & Jena, 2006). At the same time, 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 two of the well-documented practical concerns (see e.g. Kling et al. (2012)) that have been found to also apply when obtaining WTP estimates for a QALY (Ahlert et al., 2016; Bobinac et al., 2012; Gyrd-Hansen et al., 2014; Soeteman et al., 2017).

The well-being valuation approach, on the other hand, avoids some of the challenges associated with stated preferences methods by relying on observational data. Further, by using large-scale general population surveys, it promises to provide a more inclusive picture of the wide range of preferences over health and wealth across various sub-populations within a given country's society. In addition, publicly available panel data surveys would allow for a continuous re-assessment of derived estimates for subsequent years with moderate effort compared to experimental methods. However, the approach limits the scope to respondents' individual *ex-post* valuations. Furthermore, endogeneity concerns are a prevailing issue of this approach, as it relies on the estimation of causal effects of health and income to calculate their marginal trade-offs.

The study by Huang et al. (2018) was the first conceptualisation and application of the well-being valuation approach for estimating v_Q. Consequently, further exploration of the approach is needed to be able to judge whether the corresponding estimates are indeed helpful for informing v_Q, also next to WTP-based estimates. This paper, therefore, 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. This for example includes different functional form assumptions regarding the link between income and utility, the construction of health utilities, and the health state dependence of the marginal utility of consumption (see e.g. Finkelstein et al. (2013)). By using German data an additional contribution lies in providing information on v_Q for a context in which such estimates are scarce, which is a likely result of German health authorities not (explicitly) basing their reimbursement decisions on the framework outlined in Equation (1). Instead the trade-off between Δc_t and ΔQ is discussed and determined in closed-door price negotiations between health authorities and the manufacturer. Whether, and to what extent, v_Q estimates influence these negotiations or whether such estimates will become more relevant in the future due to changes in legislation is unknown.

For our analysis, we used data from the German Socio-Economic Panel (SOEP) from 2002 to 2018 providing information on a sample of 29,735 individuals followed over multiple periods. Fixed effects models 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.

2 Methods

2.1 Conceptual framework

We generally followed the framework proposed by Huang et al. (2018) for obtaining v_Q based on the well-being valuation approach. 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:

$$W_{it} = W(Y_{it}, H_{it}) \tag{2}$$

where W_{it} is a vector of the individual's well-being at all observed time points (w_{it}), Y_{it} is a vector containing 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;

$$W_i = \sum_{t=0}^{T} W(Y_{it}, H_{it}) \tag{3}$$

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

$$W_i = W(Y_i + \Delta Y_i, H_i + \Delta H_i) \tag{4}$$

The chosen approach estimates the population average ΔY necessary, to offset an imposed hypothetical change in health state H over the period T equivalent to one QALY. Therefore we refer to ΔY as the compensating income variation for one QALY, or short CIV_{QALY}.

2.2 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 is a widely used approach. The appropriateness of the required cardinality assumption of life satisfaction was shown by Ferrer-i Carbonell and van Praag (2002). The underlying empirical model takes the following form;

$$W_{irt} = \alpha + \beta_0 H_{irt} + \beta_1 H_{irt-1} + \delta_0 Y_{irt} + \delta_1 Y_{irt-1} + \tau X_{irt} + \lambda_i + \mu_r + \epsilon_t + u_{irt} \tag{5}$$

where W_{irt} refers to the subjective well-being of individual i living in region r at time t, which we captured using self-reported life satisfaction. The individual's health status H_{irt} is captured by health utility values based on the short form six dimensions (SF-6D) instrument and the original UK utility tariff for the SF-6D (Brazier & Roberts, 2004). 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 confounders, which could have affected the individual's well-being next to health and income. To account for the impact of time-invariant unobservables, we incorporated individual (λ_r), state (μ_r), and time (ϵ_t) fixed effects, with the remaining error term being u_{irt}. Heteroscedasticity-robust standard errors were used in all calculations.

In a second step the CIV_{QALY} values were obtained by dividing the estimated health status coefficients (β_0 and β_1) by the coefficient estimates of income (δ_0 and δ_1):

$$CIV_{QALY} = \frac{\beta_0 + \beta_1}{\delta_0 + \delta_1} \tag{6}$$

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.

2.3 Instrumental variable specification

A well-documented problem of the well-being valuation approach is the likely endogeneity of the income coefficient estimate. This was frequently addressed using an instrumental variable (IV) approach (see e.g. Howley (2017), McNamee and Mendolia (2018), and Brown (2015)). Huang et al. (2018) instrumented income with the occurrence of financial-worsening-events such as personal bankruptcy or large financial losses. In their analysis the differences between OLS- and IV-based coefficient estimates and the resulting CIV_{QALY} values were considerable, leading to a 130-fold differences in monetary valuations.

Lacking information on financial-worsening events we explored alternative instruments which have previously been applied with SOEP data: These instruments related to either past or future income (Bayer & Juessen, 2015; Katsaiti, 2012), or industry wage structure (Luechinger, 2009; Pischke, 2011). As our base model already included lagged income we adopted the approach developed by Luechinger (2009), who used predicted labour-market earnings based on industry-occupation cells as instrument for income.

The rationale of this instrument 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. This income variance is therefore not reflecting individual-level efforts 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 type of income instrument is that the captured income shifts have a rather permanent nature, whereas financial worsening events (as used by Huang et al. (2018)) or lottery wins are often can be highly transitory shocks. In addition permanent income shifts have been found to be of higher relevance for individuals' well-being (Bayer & Juessen, 2015; Cai & Park, 2016).

The identifying assumption is therefore that the income variance 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;

$$L_{irt} = \alpha + \rho_0 I_{irt} + \rho_1 O_{irt} + \rho_2 T_{irt} + \rho_3 R_{irt} + \mu_r + \epsilon_t + u_{irt} \tag{7}$$

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.[1] Deviating from Luechinger (2009), who predicted labour earnings for around 5,000 industry occupation cells, we followed Pischke (2011) and collapsed the number of industry branches and occupation groups to 33 and 22, respectively, forming a total of 726 industry-occupation cells. The obtained predicted labour earnings were

[1]Models were run separately for East and West Germany to account for the persisting income and labour market differences

summed on the household level and weighted by household composition to obtain the predicted household labour income $\widehat{L}^{HH}_{irt}$, the instrument used in the first-stage regression;

$$Y_{irt} = \alpha + \bar{\beta}_0 H_{irt} + \bar{\beta}_1 H_{irt-1} + \bar{\delta}_0 \widehat{L}^{HH}_{irt} + \bar{\delta}_1 \widehat{L}^{HH}_{irt-1} + \bar{\tau} X_{irt} + \bar{\lambda}_i + \bar{\mu}_r + \bar{\epsilon}_t + \bar{u}_{irt} \tag{8}$$

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

$$W_{irt} = \alpha^I + \beta^I_0 H_{irt} + \beta^I_1 H_{irt-1} + \delta^I_0 \widehat{Y}_{irt} + \delta^I_1 \widehat{Y}_{irt-1} + \tau^I X_{irt} + \lambda^I_i + \mu^I_r + \epsilon^I_t + u^I_{irt}. \tag{9}$$

The resulting coefficients for health (β^I_0 and β^I_1) and income (δ^I_0 and δ^I_1) were then included in Equation (6) to calculate the IV CIV_{QALY} estimate. All regressions were conditioned on having at least two consecutive observations per individual. Income outliers (as will be defined in section 2.4) were dropped from the base case analysis.

2.4 Alternative specifications

The following will outline our efforts to address several empirical and conceptual issues related to applying the well-being valuation method to estimate a CIV_{QALY}, which were not, or only briefly, discussed in the study by Huang et al. (2018).

Regional differences and time periods

To explore regional variation in CIV_{QALY} estimates, we separated our sample into East and West Germany, motivated by the persisting differences in life satisfaction and income levels (Frijters et al., 2004; Vatter, 2012). Temporal periods were investigated due to concerns of the (undesired) impact of national macro economic conditions on CIV_{QALY} estimates. Huang et al. (2018) reported that the chosen time periods had little effect on their CIV_{QALY} estimates. However, this may be different in our case as Germany, unlike Australia, experienced considerable economic fluctuations before and after the global financial crisis, and underwent substantial labour market reforms between 2002 to 2018, partly in response to these fluctuations.

Treatment of outliers

Due to a right-skewed and long-tailed income distribution, with self-reported income often misreported or even exaggerated (Hariri & Lassen, 2017), income outliers may have a large effect on CIV_{QALY} estimates when linear models are applied (Rousseeuw & Leroy, 1987). To identify outliers, which remains challenging for fixed effects models (Verardi & Croux, 2009), we reformulated our base case model as a pooled OLS model and calculated DFbeta, a measure of influence, which quantifies the impact that dropping an observation has on the coefficient estimate. All observations with a DFbeta larger than 1, a recommended threshold (Bollen & Jackman, 1985), were dropped from the base case analysis. In a robustness check, the calculations were repeated including the identified outliers.

Income specification

We log-transformed income to accommodate for the diminishing marginal return of income (Layard et al., 2008), and reduce the impact of outliers. CIV_{QALY} was estimated based on a slightly modified equation as used by Ólafsdóttir et al. (2020) and van den Berg and Ferrer-i Carbonell (2007). This entailed dropping the lagged income and health coefficients as used our base model (Equation 6).

$$CIV_{QALY} = \overline{y} * \left(exp\left(\frac{-\beta_0 * \frac{1}{\Delta}}{\delta_0} \right) - 1 \right) * \Delta \tag{10}$$

In the log-income specification, CIV_{QALY} was calculated as the percentage share of yearly population income (here yearly median income $\overline{y}$). By construction CIV_{QALY} values would be confined to be no greater than the income level which may be acceptable when valuing small gains or changes but not when valuing 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 change in 0.1 QALYs 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 piece-wise linear specification similar to Ólafsdóttir et al. (2020). To obtain the appropriate number of income splines and cut-off values, an iterative process, starting with the ten deciles as cut-offs, was chosen. The equality of coefficient estimates of adjacent splines was tested and non-significantly different splines were gradually combined until all coefficients were significantly different and model fit did not improve. CIV_{QALY} values were then calculated for each income spline separately, 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 SF-6D value set, calculated using time-trad-off tasks (Brazier & Roberts, 2004), to construct health utilities. 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 (Jonker et al., 2018).

Health state dependence of utility of consumption

Another empirical issue of concern relates to the interaction between health and income with regards to its impact on experienced (consumption) utility. This so-called health state dependence implies that the marginal gain in experienced utility from a given income change is directly dependent on an individual's underlying health status (Finkelstein et al., 2013). So far, there is only scarce and inconclusive evidence on the magnitude and the direction of this effect: Finkelstein et al. (2013) found a negative health state dependence based on US data, i.e. a higher marginal utility of income in good health compared to bad health. However, replicating their approach based on European data, Kools and Knoef (2019) found evidence for positive health state dependence, potentially due to differing institutional environments impacting the provision of public goods, and the more generous European healthcare systems.

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 good and bad health states. Finkelstein et al. (2013) used the onset of chronic diseases for this purpose. 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 younger and healthier groups. First, we reduced the sample to those 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.[2] This was done to ensure that individuals in this group have experienced a consequential change in their mental 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 and 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 A3 for additional details). Importantly, the sample included individuals transitioning from good to bad health and vice versa, although the transition from good to bad is the most frequently observed.

[2]The SF-12 is also used to calculate SF-6D health utilities. Mental and physical component scores range from 0 (worst) to 100 (best) with a normalised mean of 50 and standard deviation of 10 (Ware et al., 1995).

3 Data

We used data from the Socio-Economic Panel (2019), or SOEP, an annually conducted large scale longitudinal survey of a representative sample of the adult (aged 16+) German population (Goebel et al., 2019). SF-6D health utilities were constructed from SF-12 data, a generic measure of health status, which is biennially included in the SOEP survey since 2002. The original utility tariff of the SF-6D for the UK (Brazier & Roberts, 2004) was applied in the absence of a Germany-specific tariff. 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 the intermediate years. However, this was only done if individuals were observed for three consecutive years and biannually provided full SF-12 data.

Life satisfaction was measured using responses to the question *"How satisfied are you with your life, all things considered?"* 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 Hagenaars et al. (1994). This entailed assigning a weight of 1 to the first adult, 0.5 to each additional adult, and 0.3 to children below the age of 16 living in the household. Income data was converted to 2018 prices using the official consumer price indices published by the Federal Statistical Office of Germany.[3]

To construct our instrument, predicted household labour income, we extracted information on net labour income and individuals' industry and occupation. Households with individuals, where information on labour income, but not on industry/occupation was available, were dropped (11,471 individuals). Predicted labour income was assumed to be zero for all individuals with no labour income information or who stated that they were not employed, to prevent dropping a considerable part of our observations.[4]

We furthermore gathered information on a similar set of variables as used by Huang et al. (2018) to control for confounding factors. These included age, disability status, marital status, educational attainment, time spent on leisure activities, and employment status.

Table 1 summarises the key characteristics of the analysis sample, consisting of 29,735 individ-

[3] Annual consumer price indices can be downloaded from the GENESIS Online Data Repository All results are based on annual CPI rates released in February 2019.

[4] Following Luechinger (2009) we added a constant of 1 to all incomes for the log-income specification.

uals providing a total of 186,906 individual-year observations.[5] As the exclusion of individuals without at least two consecutive SF-6D values was the only major exclusion criterion, the analysis remains largely representative for the overall population of Germany. Over the period between 2002 and 2018, mean life satisfaction was 7.09 (1.71), and mean net monthly equivalised household income was €2,029 (SD 1,29). Applying the SF-6D scoring algorithm produced health utilities with a mean of 0.73 (SD 0.13).

Table 1: Descriptive statistics

Variable	**Mean**	**Std. Dev.**	**Description**
Life satisfaction	7.09	1.71	0 (lowest) to 10 (highest)
Income in 1000's	2.03	1.29	Monthly household income in €
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: Tertiary	0.63	0.48	1 if secondary educated
Education: Secondary	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	

[5]Appendix Table A1 provides an overview of the conditioning applied to the SOEP data.

4 Results

4.1 Baseline results

The baseline OLS and IV fixed results are shown in Table 2. We were able to predict labour income for 20,618 individuals yielding 116,125 observations. The instrument passed the Cragg-Donald weak identification test (F-value: 1,863.7) and the Kleibergen-Paap underidentification test (χ^2: 3,642.0), indicating a high relevance of the instrument as is common with such income-based instruments (Bayer & Juessen, 2015; Luechinger, 2009). 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 model. 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 (Bayer & Juessen, 2015; Pischke, 2011). Interestingly, the difference is minimal compared to what was observed by Huang et al. (2018), who reported an IV coefficient which was 130 times larger than the OLS coefficient (0.080 and 0.0006).

Table 2: Baseline results

	OLS		IV	
Income in 1000's	0.048***	(0.005)	0.098***	(0.032)
Income in 1000's $(t-1)$	0.007	(0.005)	0.043	(0.027)
SF-6D utility	3.121***	(0.064)	3.115***	(0.054)
SF-6D utility $(t-1)$	0.104*	(0.060)	0.098*	(0.054)
Disability	-0.138***	(0.022)	-0.137***	(0.017)
Age	0.093***	(0.014)	0.084***	(0.015)
Age squared	-0.000***	(0.000)	-0.000**	(0.000)
(de facto) Married	0.183***	(0.023)	0.176***	(0.016)
Primary education	-0.184*	(0.095)	-0.210***	(0.077)
Tertiary education	-0.180***	(0.056)	-0.190***	(0.048)
Leisure time	0.031***	(0.005)	0.030***	(0.004)
Leisure time squared	-0.002***	(0.001)	-0.002***	(0.000)
Unemployed	-0.525***	(0.028)	-0.529***	(0.020)
Work hours	0.002***	(0.000)	0.001***	(0.000)
Tenure	-0.006***	(0.001)	-0.007***	(0.001)
Individuals * Years	186,902		186,902	
Individuals	29,735		29,735	
Model statistics				
Cragg-Donald Wald F statistics			1,863.7	
Anderson canon. corr. LM statistics			3,642.0	
Endogeneity test			10.0	
BIC	540,755		540,995	
CIV in €	58,533		22,717	

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

Applying the estimated income and SF-6D coefficients to Equation (6) resulted in a CIV_{QALY} value of €58,533 in the OLS model. This value represents the average amount of additional income necessary to maintain the same level of life satisfaction if a hypothetical health change of 1 QALY is imposed. The corresponding value for the IV estimates was €22,717.

Table 3 columns 2-3 contains estimates for East and West Germany separately. 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, respectively), 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. This difference may be explained by the prevailing differences in (household) income between West and East. While the average monthly equivalised income in the sample was €2,140 in the West, it was only €1,652 in the East. Considering a diminishing marginal return, income changes consequently have a higher impact on life satisfaction in the East.

As shown in Table 4 (columns 4-6), excluding the years of the financial crisis and recession in Germany (2007-2009) had only a minor impact on the OLS and IV CIV_{QALY} values (€54,567 and €20,574, respectively). 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 2010-2018 (€70,572 and €24,811). This resulted from larger estimated effects of income on life satisfaction in earlier periods, which may both be a result of the generally positive income development or a shift in population preferences and values over the last decades. Appendix Table A2 provides further results on age and gender subgroups.

Table 3: Results by region and time-period

	Baseline		**East**		**West**		**w/o 2007-2009**		**2002-2006**		**2010-2018**	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Income in 1000's	0.05***	0.10***	0.13***	0.18**	0.04***	0.07**	0.05***	0.11***	0.06***	0.29***	0.04***	0.09*
	(0.01)	(0.03)	(0.02)	(0.08)	(0.01)	(0.04)	(0.01)	(0.04)	(0.01)	(0.09)	(0.01)	(0.05)
Income in 1000's $(t-1)$	0.01	0.04	0.00	0.03	0.01	0.04	0.01*	0.05	-0.00	0.10	0.01	0.04
	(0.01)	(0.03)	(0.02)	(0.06)	(0.01)	(0.03)	(0.01)	(0.03)	(0.01)	(0.08)	(0.01)	(0.04)
SF-6D utility	3.12***	3.12***	2.90***	2.90***	3.18***	3.17***	3.16***	3.15***	2.93***	2.92***	3.08***	3.08***
	(0.06)	(0.05)	(0.13)	(0.12)	(0.07)	(0.07)	(0.07)	(0.07)	(0.15)	(0.15)	(0.08)	(0.08)
SF-6D utility $(t-1)$	0.10*	0.10*	-0.12	-0.12	0.16**	0.16**	0.10	0.09	0.06	0.06	-0.07	-0.07
	(0.06)	(0.05)	(0.12)	(0.12)	(0.07)	(0.07)	(0.07)	(0.07)	(0.14)	(0.14)	(0.08)	(0.08)
Model statistics												
Cragg-Donald		1,863.7		323.9		680.2		783.4		181.2		494.3
Anderson		3,642.0		544.4		1,265.5		1,429.5		328.8		907.3
Endogeneity test		10.0		1.5		5.8		9.7		8.2		2.7
BIC	540,755	540,995	127,072	127,092	412,723	412,877	431,238	431,487	129,869	130,432	276,374	276,464
Observations	186,902	186,902	43,447	43,447	143,361	143,361	151,461	151,461	48,678	48,678	101,048	101,048
CIV in €	58,533	22,717	20,750	12,982	75,748	28,548	54,567	20,574	56,640	7,720	70,572	24,811

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

4.2 The impact of income specification

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 4 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 observations points (1 year) with life satisfaction remaining constant at 10.

In the models using log-transformed income (Table 4 columns 5-6), the income coefficient was 0.24, larger than reported before using the SOEP by Pischke (2011) (0.125 to 0.182). The corresponding IV coefficient, with a value of 0.63, was close to previous IV estimates based on the industry-wage structure and the SOEP: Luechinger (2009) reported an estimate of 0.55, while Pischke (2011) reported values ranging from 0.489 to 0.617 across specifications. Previous estimates based on instruments using lagged or future income shocks were also similar, with Katsaiti (2012) reporting coefficients ranging from 0.323 to 0.4557 and Bayer and Juessen (2015) providing a range of 0.45 to 0.50 for the impact of permanent income shocks.[6] Compared to our baseline, the log transformation resulted in considerably larger CIV_{QALY} values. 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.[7]

[6]Bayer and Juessen (2015) omitted East Germany from their analysis, which may have lead to a downward bias in their income coefficients due to the overall higher income levels in West Germany.

[7]Huang et al. (2018) did not observe such considerable differences between linear and log income based estimates. However, they multiplied the ratio of income and health coefficients as in Equation (6) with the median income (as opposed to Equation (10)). Applying this to our data resulted in even larger CIV_{QALY} estimates of €256,210 (OLS) and €116,620 (IV).

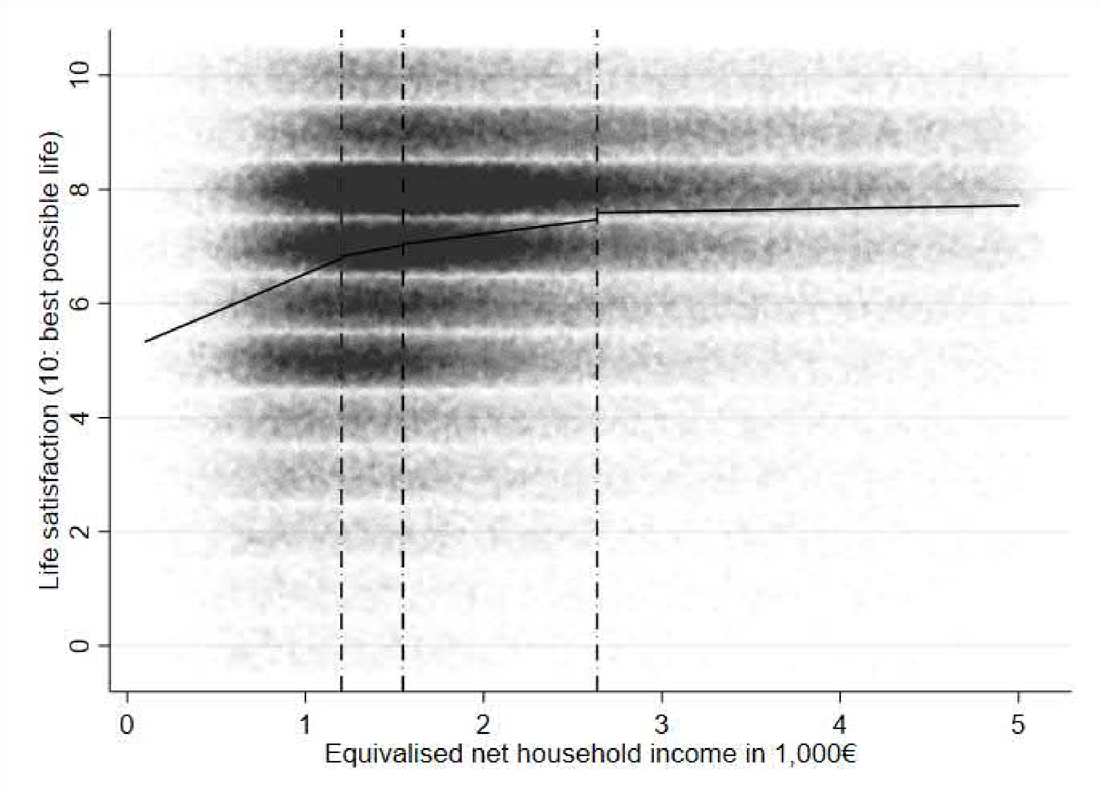
Table 4: Income specifications

	Baseline		**With Outliers**		**Log income**		**Piecewise**
	OLS	IV	OLS	IV	OLS	IV	OLS
Income in 1000's	0.05***	0.10***	0.03***	0.10***			
	(0.01)	(0.03)	(0.01)	(0.03)			
Income in 1000's $(t-1)$	0.01	0.04	0.01***	0.04			
	(0.01)	(0.03)	(0.00)	(0.03)			
SF-6D utility	3.12***	3.12***	3.12***	3.12***	3.18***	3.16***	3.18***
	(0.06)	(0.05)	(0.06)	(0.06)	(0.05)	(0.05)	(0.05)
SF-6D utility $(t-1)$	0.10*	0.10*	0.10*	0.10*			
	(0.06)	(0.05)	(0.06)	(0.06)			
Log income					0.24***	0.63***	
					(0.02)	(0.13)	
1^{st} income spline							0.43***
							(0.05)
2^{nd} income spline							0.27***
							(0.05)
3^{rd} income spline							0.11***
							(0.02)
4^{th} 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 in €	58,533	22,717	82,484	22,782	153,877	81,649	97,486
w/o 4^{th} spline							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 1 plots the overall distribution of life-satisfaction across income, and the linear fit of life satisfaction across income splines. The coefficients of the four income splines in the piece-wise regression were 0.43, 0.27, 0.11, and 0.01, depicting a non-linear, diminishing pattern. The corresponding CIV_{QALY} values for the four income splines were €7,347, €11,686, €29,548 and €409,810, respectively. 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 non-significant. Just using the lower three splines lead to a CIV_{QALY} value of €19,515.

Figure 1: Relationship between life satisfaction and income across income splines



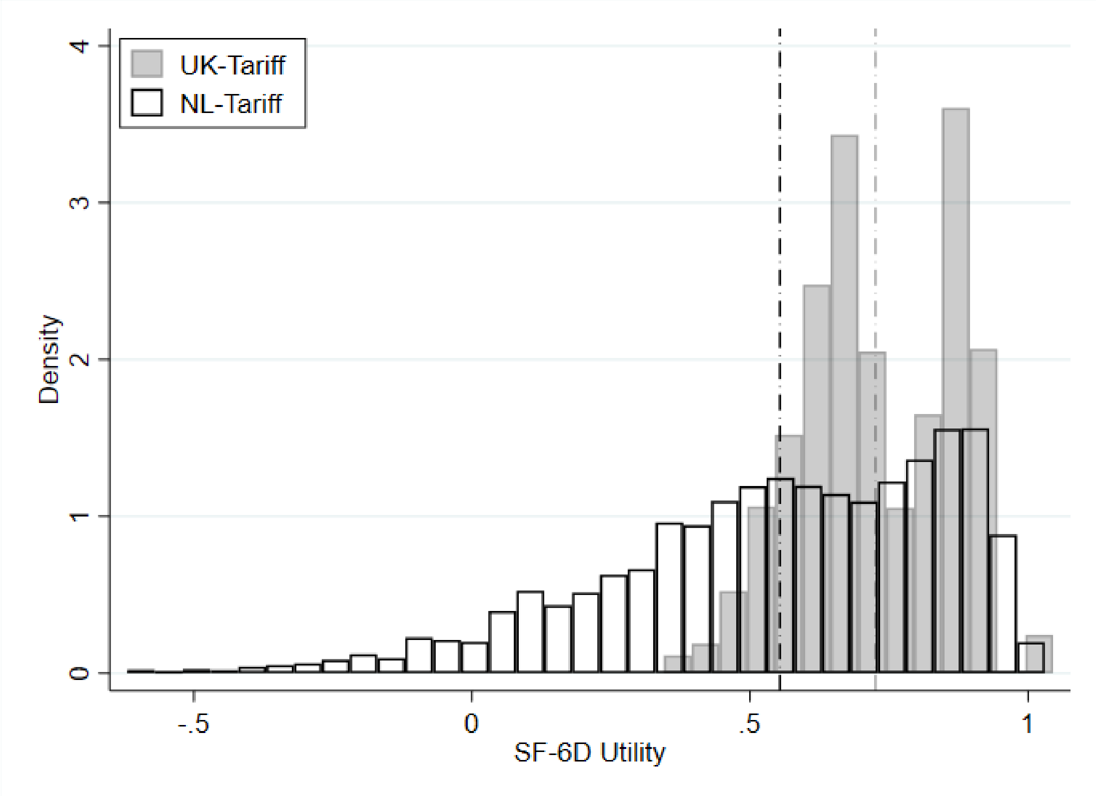
Note: Life satisfaction values are depicted as small grey dots. Black dash-dotted vertical lines represent the income splines used in the piece-wise linear regression. Black horizontal lines plot the linear fit within these splines.

4.3 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 2), with the mean SF-6D utility decreasing from 0.725 to 0.554. These differences may more likely reflect methodological differences 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 (Norman et al., 2009).

Figure 2: SF12 index values using UK and Dutch tariffs



Note: 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 5. 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). This decrease resulted 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 5: Choice of SF-6D tariffs

	UK Tariff		Dutch Tariff	
	OLS	IV	OLS	IV
Income in 1000's	0.05***	0.10***	0.05***	0.09***
	(0.01)	(0.03)	(0.01)	(0.03)
Income in 1000's $(t-1)$	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		907.1
Anderson		3,642.0		1,671.4
Endogeneity test		10.0		9.4
BIC	540,755	540,995	538,297	538,523
Observations	186,902	186,902	186,902	186,902
CIV in €	58,533	22,717	32,534	13,054

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 (see section 2.4). 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). Table 6 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 status" observations, the corresponding CIV_{QALY} estimates were lower with € 33,336 and € 16,532. For "bad health status", the OLS-based CIV_{QALY} estimate was € 38,374 and the IV-based estimate € 11,779.

Table 6: Health state dependence

	Baseline		Good Health		Bad Health	
	OLS	IV	OLS	IV	OLS	IV
Income in 1000's	0.07***	0.17**	0.05***	0.11	0.08**	0.32
	(0.01)	(0.07)	(0.02)	(0.08)	(0.04)	(0.24)
Income in 1000's $(t-1)$	0.03**	0.02	0.03**	0.05	0.03	0.05
	(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 $(t-1)$	0.10	0.11	0.12	0.12	0.32	0.32
	(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 in €	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 considerable 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 also 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). Unfortunately, we were not able to follow Kools and Knoef (2019) and Finkelstein et al. (2013) in focusing on non-working individuals, which would have ensured stable income across health states, ruling out that the increase in the income coefficients was driven by individuals losing their income, and hence having a larger marginal utility of additional earnings. For our analysis, such a restriction was not feasible, as within-person income variation is necessary to estimate the income coefficients in fixed-effects models. However, the general empirical pattern remains the same when excluding individuals with large negative income differences between good and bad health states (see Appendix Table A5). This also holds when further restricting the sample to only the working population (Table A6) and those experiencing severe health changes (Table

A7).

4.4 Robustness checks

Lastly, we tested the robustness of our baseline results to some general concerns regarding our estimation strategy (Table 7). First, we re-estimated the OLS and IV models without imputing SF-6D utilities for the years where SF-12 data was not collected. This resulted in a sample of 85,433 observations across 21,718 individuals. The resulting CIV_{QALY} estimates based on 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 small differences were driven by larger SF-6D coefficients compared to the baseline calculations. This effect likely was a result of smoothing health utility changes, by linearly imputing between years, and therefore reducing the within-person variance of health status.

In a second robustness check, we limited our sample to individuals which were in paid employment and provided industry-occupation information. This is 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 the working population being slightly healthier as individuals without labour income mainly due to the age difference. The sum of both income coefficients was smaller in the corresponding IV-calculations compared to baseline, shifting the CIV_{QALY} upwards.

Lastly, we followed Luechinger (2009) by excluding households in which the main income earner was self-employed. The reasoning behind this robustness check was that among these individuals, the income measurement error was likely to be amplified. Self-employed individuals are often reluctant to disclose their true income while also experiencing less stable income streams and hence, even if not reluctant to report, they might simply misreport by mistake. The resulting CIV_{QALY} estimates and income and SF-6D coefficients were similar to the baseline estimates (€55,359 and €20,352).

Table 7: Robustness checks

	Baseline		**No Imputation**		**Working only**		**no Self-Employed**	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Income in 1000's	0.05***	0.10***	0.05***	0.14***	0.05***	0.05	0.05***	0.10***
	(0.01)	(0.03)	(0.01)	(0.05)	(0.01)	(0.03)	(0.01)	(0.04)
Income in 1000's $(t-1)$	0.01	0.04	-0.00	-0.00	0.01	0.07**	0.00	0.06**
	(0.01)	(0.03)	(0.01)	(0.07)	(0.01)	(0.03)	(0.01)	(0.03)
SF-6D utility	3.12***	3.12***	3.52***	3.51***	2.95***	2.94***	3.14***	3.14***
	(0.06)	(0.05)	(0.06)	(0.05)	(0.08)	(0.07)	(0.07)	(0.06)
SF-6D utility $(t-1)$	0.10*	0.10*	0.47***	0.46***	0.07	0.06	0.08	0.08
	(0.06)	(0.05)	(0.05)	(0.05)	(0.07)	(0.07)	(0.06)	(0.06)
Model statistics								
Cragg-Donald		1,863.7		192.1		1,355.7		2,239.4
Anderson		3,642.0		382.2		2,637.7		4,345.1
Endogeneity test		10.0		5.8		5.4		11.8
BIC	540,755	540,995	236,338	236,538	319,169	319,323	499,342	499,565
Observations	186,902	186,902	85,433	85,433	116,125	116,125	172,031	172,031
CIV in €	58,533	22,717	80,522	28,130	52,829	26,097	55,359	20,352

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

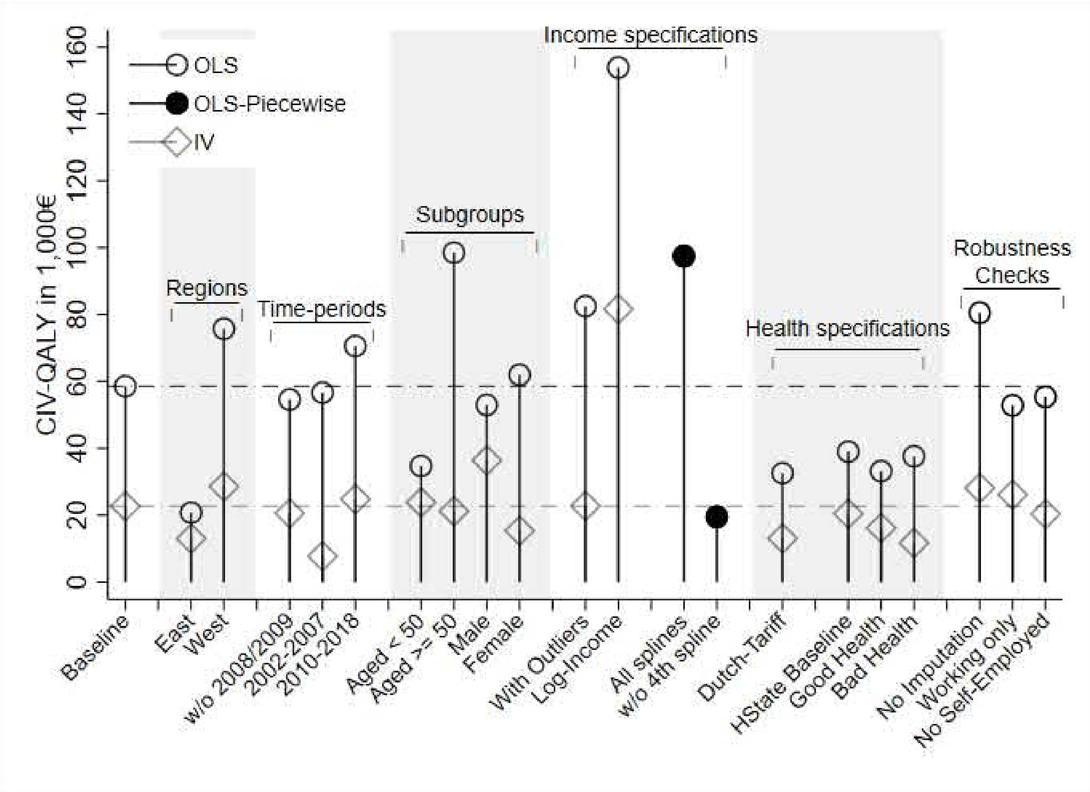
5 Discussion

While estimates of the monetary value of a QALY (v_Q) historically were mainly based on stated preference WTP experiments, we used an alternative strategy previously applied by Huang et al. (2018), utilising large-scale observational data from Germany: the well-being valuation approach. Beyond demonstrating the general feasibility of this method in a different country context we explored several empirical and methodological challenges with important consequences for the practical usefulness of well-being valuation based v_Q estimates (CIV_{QALY}), and provide estimates of v_Q for Germany.

5.1 Overview and context of results

Figure 3 presents an overview of the estimated CIV_{QALY} values across subgroups and specifications. The baseline calculations provided average monetary valuations of a QALY of €58,533 and, when instrumenting for income, €22,717. The corresponding estimates in Huang et al. (2018) for Australia were €2,149,324 and €45,586. Our 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 (e.g. Ólafsdóttir et al. (2020)), with the IV estimates remaining rather stable around €20,000 per QALY. The range of CIV_{QALY} estimates obtained in our study fit into the ballpark of more reasonable stated preference estimates (Ryen & Svensson, 2015). Furthermore, it is important to note that all of the 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). Their stated preference based estimates constituted the only v_Q estimates for Germany up until now. A first approximation of an *opportunity cost based QALY threshold value*, or k_Q, for Germany was reported by Woods et al. (2016). Using empirical estimates of health care opportunity costs for the UK, 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).

Figure 3: Overview of CIV_{QALY} estimates



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

5.2 Limitations and strengths of the analysis

We have to acknowledge several limitations of our analysis, first and foremost relating to the instrumental variable approach. IV-based estimates rely on a set of restrictive assumptions, relating to both their unbiasedness and their general informational value. A valid concern is that occupational choice may be related to other unobserved confounders, such as personality traits or individual preferences over income (Pischke & Schwandt, 2012). The use of individual fixed effects should somewhat alleviate concerns related to this due to the rather stable nature of personality traits (Borghans et al., 2008) but naturally they cannot provide complete assurance. 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 (Angrist et al., 1996). Using predicted labour income as an instrument, at least questions the generalisability of our IV estimates to the full, also non-working, population.

Important to note is that income variation in industry-occupation cells predominantly consists of *positive, upward shifts in wages* (and differences therein). This is conceptually different to using financial worsening events as income instrument, as done by Huang et al. (2018),[8] as their instrument captures the impact of *income losses*. Given that robust findings from behavioural economics showing the utility impact of losses to be higher than the impact of similar gains (see for example Attema et al. (2016) for the case of health states), our IV based CIV_{QALY} estimates likely represent a lower-bound.

The potential endogeneity of health (status) in life satisfaction regressions, which is rarely addressed in the related literature, is a further concern, given evidence hinting at a reverse causal relationship (see e.g. Veenhoven (2008) or Sabatini (2014)). Endogeneity could be addressed by finding an appropriate instrument for health or identifying plausibly exogenous health shocks. However, this is not straightforward in practice and it is questionable how generalisable such localised causal effects would be for the overall effect of the multi-dimensional construct of health on life satisfaction. More practical limitations, which we explored above, were that we linearly impute SF-6D utilities for every second year to make full use of the SOEPs rich annual data. The imputation required us to condition the sample on individuals who had at least three consecutive observations. This may have resulted in underestimating the impact of deteriorating health, as 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, since subjective well-being enters the model twice: First, as an implicit consideration in the SF-6D health state valuation tasks, and secondly, as proxy for experienced utility (Equation 2). 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, which was rescaled to a 0 to 1 range (artificially 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 A3, 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.

[8] Ambrosio et al. (2018) found direct and long-term impact of financial worsening (and improvement) events on life satisfaction and health behaviours beyond income-changes using the HILDA dataset, raising concerns on the general appropriateness of using such events as income instruments.

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 a priori 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 A3). While these sensitivity checks somewhat alleviate the concerns about double-counting, the latter revealed that a considerable part (46 percent) of the cumulative CIV_{QALY} value stemmed from the large impact of the mental health dimension 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 utility?

One strength of our study is that we provided additional evidence on the applicability of the well-being valuation approach in the context of estimating v_Q empirically, precisely to highlight its limitations and so guide future research and stir an open debate about this approach. We addressed several important conceptual and empirical issues, provided a thorough discussion of the limitations, and suggest possible solutions for some of them. A further strength of our study is that we could base our analysis on a large-scale, long-running, and extensive dataset, allowing us to explore the impact of a wide array of specification choices. Lastly, despite the mentioned issues, most estimated v_Q values had clear face validity when compared to existing values (however determined) for Germany and neighbouring countries (Ahlert et al., 2016; Ryen & Svensson, 2015).

5.3 Implications for future applications of the approach

There are several practical implications of this 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 4), 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 possible values has a large impact (Table A3), 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. While it is convenient to opt for a country tariff whose origin can be placed in cultural and socio-economic proximity to the country to be investigated the impact of methodological peculiarities in how these tariffs were generated should not be ignored. Further, if competing tariffs are available results should be provided for the alternatives. On a side note, 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, the EQ-5D is not routinely included in datasets such as the SOEP. Lastly, the differing values obtained when considering East and West Germany separately, or specific time periods (Table 3), also highlight the potential importance of country-context and macroeconomic conditions for CIV 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 the marginal utility of consumption. 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 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 6). 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. Hence, a pragmatist might argue that this issue is of theoretical interest only. We would argue, however, that this is an inherent limitation of 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. For observational data, there seems to be no readily available solution, although access to administrative health records would allow for a better assessment of the scope of this blind spot in the data.

An additional conceptual concern related to health state dependence is the question of adaptation to bad health over time (Huang et al., 2018). This hedonic adaptation implies the gradual return of subjective well-being to pre-health-shock levels despite continued (or deteriorating) bad health (Loewenstein & Ubel, 2008). This phenomenon has been documented before using the SOEP-data (Oswald & Powdthavee, 2008) and would generally decrease estimated CIV_{QALY} as the marginal utility with respect to health would decrease over time spend 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 should be considered at all when quantifying this impact. The recent findings by Etilé et al. (2020), who documented a heterogeneous distribution of adaptive potential across subgroups, underline the potential relevance this conceptual 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 exploiting natural experiments or setting up experiments similar to the basic income experiment in Finland (Kangas et al., 2019). If valid and stable estimates can be found, these could at least serve as (national) reference values, and could be used in robustness checks for specific well-being valuation studies to test the external validity of estimates. Ideally, one would also see the regular inclusion of variables that represent valid instruments for income into general population panel surveys, therefore allowing for cross-national replications of results. This would greatly increase the possibility to explore the reliability and validity of the well-being valuation approach to valuing QALYs across countries. Meanwhile, future applications may draw upon recent advances into the generalisability of IV-based estimates (see e.g. Mogstad et al. (2018)) to explore how these concerns can be addressed using these methods.

6 Conclusions

Our study confirms that estimating the value of a QALY based on the well-being valuation approach and large-scale observational data is feasible and leads to plausible v_Q estimates for Germany. Health care funding decisions in Germany are currently not reliant on cost utility analysis or a v_Q based threshold value, at least in part because defining such a threshold was considered to be too difficult (Bundesministerium für Gesundheit, 2008). Finding monetary estimates using a compensating income variation approach that are in the same ballpark as those based on stated preference studies to some extent puts this into perspective. Whether, and in which direction this influences the German discussion and contributes to Germany adopting a more explicit and transparent health care decision-making framework is unclear.

While we showed that some empirical and conceptual challenges of applying the well-being valuation approach for estimating v_Q may have limited impact on estimates and may be easily overcome, other issues will remain challenging for future applications of the approach for valuing QALYs (or health in general). Future researchers could address these challenges further, but may also reveal additional limitations. At the same time, further exploring the validity of alternative approaches of estimating v_Q is necessary. Stated preference WTP experiments or methods aimed at eliciting the value of a statistical life, as recently done by Herrera-Araujo et al. (2020), continue to provide important insights into the empirical estimation of v_Q.

These different approaches to estimating the value of health have their unique advantages and disadvantages while providing conceptually different v_Q estimates. Given their complementary strengths and limitations methodological diversity is desired in the ongoing endeavour of measuring the monetary value of health. The importance of obtaining such values has rarely been as obvious as during the current pandemic. Governments around the globe have to decide about drastic and intrusive countermeasures to prevent the spread of a virus to avoid the associated morbidity and mortality while facing substantial social and economic costs. Estimates of the public's monetary valuation of health are crucial for informing uncomfortable trade-offs that societies face now and in the future, both within health care but also beyond (Chilton et al., 2020).

References

Ahlert, M., Breyer, F., & Schwettmann, L. (2016, feb). How you ask is what you get: Framing effects in willingness-to-pay for a QALY. *Social Science & Medicine*, *150*, 40–48. doi: 10.1016/J.SOCSCIMED.2015.11.055

Ambrosio, C. D., Clark, A., & Zhu, R. (2018). 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 . *Working paper*. Retrieved from `http://www.iariw.org/copenhagen/dambrosio.pdf`

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, *91*(434), 444–455. doi: 10.2307/2291629

Attema, A. E., Brouwer, W. B., L'Haridon, O., & Pinto, J. L. (2016, jul). An elicitation of utility for quality of life under prospect theory. *Journal of Health Economics*, *48*, 121–134. doi: 10.1016/j.jhealeco.2016.04.002

Bayer, C., & Juessen, F. (2015). Happiness and the persistence of income shocks. *American Economic Journal: Macroeconomics*, *7*(4), 160–187. doi: 10.1257/mac.20120163

Bobinac, A., van Exel, N. J. A., Rutten, F. F., & Brouwer, W. B. (2012). Get more, pay more? an elaborate test of construct validity of willingness to pay per qaly estimates obtained through contingent valuation. *Journal of health economics*, *31*(1), 158–168. doi: 10.1016/j.jhealeco.2011.09.004

Bobinac, A., van Exel, N. J. A., Rutten, F. F. H., & Brouwer, W. B. F. (2013, oct). Valuing QALY gains by applying a societal perspective. *Health Economics*, *22*(10), 1272–1281. doi: 10.1002/hec.2879

Bollen, K. A., & Jackman, R. W. (1985, may). Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases. *Sociological Methods & Research*, *13*(4), 510–542. doi: 10.1177/0049124185013004004

Borghans, L., Duckworth, A. L., Heckman, J. J., & ter Weel, B. (2008). The Economics and Psychology of Personality Traits. *Journal of Human Resources*, *43*(4), 972–1059. doi: 10.3368/jhr.43.4.972

Brazier, J. E., & Roberts, J. (2004). The Estimation of a Preference-Based Measure of Health From the SF-12. *Medical care*, *42*(9), 851–9.

Brouwer, W., van Baal, P., van Exel, J., & Versteegh, M. (2019). When is it too expensive?

Cost-effectiveness thresholds and health care decision-making. *European Journal of Health Economics*, *20*(2), 175–180. doi: 10.1007/s10198-018-1000-4

Brown, T. T. (2015). The Subjective Well-Being Method of Valuation: An Application to General Health Status. *Health Services Research*, *50*(6), 1996–2018. doi: 10.1111/1475-6773.12294

Bundesministerium für Gesundheit. (2008). *Stellungnahme zur Methodik der Kosten-Nutzen-Bewertung von Arzneimitteln.* Retrieved from http://www.bmg.bund.de/

Cai, S., & Park, A. (2016). Permanent income and subjective well-being. *Journal of Economic Behavior and Organization*, *130*, 298–319. doi: 10.1016/j.jebo.2016.07.016

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

Chilton, S., Nielsen, J. S., & Wildman, J. (2020). Beyond covid-19: How the 'dismal science'can prepare us for the future. *Health Economics*. doi: 10.1002/hec.4114

Cleemput, I., Neyt, M., Thiry, N., De Laet, C., & Leys, M. (2011, jan). Using threshold values for cost per quality-adjusted life-year gained in healthcare decisions. *International Journal of Technology Assessment in Health Care*, *27*(1), 71–76. doi: 10.1017/S0266462310001194

de Meijer, C., Wouterse, B., Polder, J., & Koopmanschap, M. (2013). The effect of population aging on health expenditure growth: A critical review. *European Journal of Ageing*, *10*(4), 353–361. doi: 10.1007/s10433-013-0280-x

Dolan, P., Kavetsos, G., Krekel, C., Mavridis, D., Metcalfe, R., Senik, C., ... Ziebarth, N. R. (2019). Quantifying the intangible impact of the olympics using subjective well-being data. *Journal of Public Economics*, *177*, 104043. doi: 10.1016/j.jpubeco.2019.07.002

Dolan, P., & Olsen, J. A. (2001). Equity in health: the importance of different health streams. *Journal of Health Economics*, *20*(5), 823–834. doi: 10.1016/s0167-6296(01)00095-9

Etilé, F., Frijters, P., Johnston, D. W., & Shields, M. (2020). Psychological resilience to major socioeconomic life events. *IZA Discussion Paper*, *No. 13063*.

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

Finkelstein, A., Luttmer, E. F., & Notowidigdo, M. J. (2013). What good is wealth without health? the effect of health on the marginal utility of consumption. *Journal* *of the European Economic Association*, *11*(suppl_1), 221–258. doi: 10.1111/j.1542-4774.2012.01101.x

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

Frijters, P., Haisken-DeNew, J. P., & Shields, M. A. (2004). Money does matter! evidence from increasing real income and life satisfaction in east germany following reunification. *American Economic Review*, *94*(3), 730–740. doi: 10.1257/0002828041464551

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The german socio-economic panel (soep). *Jahrbücher für Nationalökonomie und Statistik*, *239*(2), 345–360.

Gyrd-Hansen, D. (2003). Willingness to pay for a qaly. *Health economics*, *12*(12), 1049–1060. doi: 10.1002/hec.799

Gyrd-Hansen, D., Jensen, M. L., & Kjaer, T. (2014). Framing the willingness-to-pay question: impact on response patterns and mean willingness to pay. *Health Economics*, *23*(5), 550–563. doi: 10.1002/hec.2932

Hagenaars, A. J., De Vos, K., Asghar Zaidi, M., et al. (1994). *Poverty statistics in the late 1980s: Research based on micro-data.* Office for Official Publications of the European Communities.

Hariri, J. G., & Lassen, D. D. (2017). Income and Outcomes: Social Desirability Bias Distorts Measurements of the Relationship between Income and Political Behavior. *Public Opinion Quarterly*, *81*(2), 564–576. doi: 10.1093/poq/nfw044

Herrera-Araujo, D., Hammitt, J. K., & Rheinberger, C. M. (2020). Theoretical bounds on the value of improved health. *Journal of Health Economics*, 102341. doi: 10.1016/j.jhealeco.2020.102341

Howley, P. (2017). Less money or better health? Evaluating individual's willingness to make trade-offs using life satisfaction data. *Journal of Economic Behavior and Organization*, *135*, 53–65. doi: 10.1016/j.jebo.2017.01.010

Huang, L., Frijters, P., Dalziel, K., & Clarke, P. (2018). Life satisfaction , QALYs , and the monetary value of health. *Social Science & Medicine*, *211*(June), 131–136. doi: 10.1016/j.socscimed.2018.06.009

Jonker, M. F., Donkers, B., de Bekker-Grob, E. W., & Stolk, E. A. (2018). Advocating a

Paradigm Shift in Health-State Valuations: The Estimation of Time-Preference Corrected QALY Tariffs. *Value in Health*, *21*(8), 993–1001. doi: 10.1016/j.jval.2018.01.016

Kangas, O., Jauhiainen, S., Simanainen, M., & Ylikännö, M. (2019). The basic income experiment 2017–2018 in Finland. Preliminary results. *Reports and memorandums of the ministry of social affairs and health*(9), 34. Retrieved from http://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161361/Report{_}TheBasicIncomeExperiment20172018inFinland.pdf{%}0Ahttp://urn.fi/URN:ISBN:978-952-00-4035-2

Katsaiti, M. S. (2012). Obesity and happiness. *Applied Economics*, *44*(31), 4101–4114. doi: 10.1080/00036846.2011.587779

Kling, C. L., Phaneuf, D. J., & Zhao, J. (2012). From Exxon to BP: Has some number become better than no number. *Journal of Economic Perspectives*, *26*(4), 3–26. doi: 10.1257/jep.26.4.3

Kools, L., & Knoef, M. (2019). Health and consumption preferences; estimating the health state dependence of utility using equivalence scales. *European Economic Review*, *113*, 46–62. doi: 10.1016/j.euroecorev.2018.12.007

Layard, R., Nickell, S., & Mayraz, G. (2008). The marginal utility of income. *Journal of Public Economics*, *92*(8-9), 1846–1857. doi: 10.1016/j.jpubeco.2008.01.007

Loewenstein, G., & Ubel, P. A. (2008). Hedonic adaptation and the role of decision and experience utility in public policy. *Journal of Public Economics*, *92*(8-9), 1795–1810.

Luechinger, S. (2009). Valuing Air Quality Using the Life Satisfaction Approach. *The Economic Journal*, *119*, 482–515. doi: 10.1111/j.1468-0297.2008.02241.x

Luechinger, S., & Raschky, P. A. (2009). Valuing flood disasters using the life satisfaction approach. *Journal of Public Economics*, *93*(3-4), 620–633. doi: 10.1016/j.jpubeco.2008.10.003

Mcdonald, R., & Powdthavee, N. (2018). The Shadow Prices of Voluntary Caregiving : Using Panel Data of Well-Being to Estimate the Cost of Informal Care. *IZA DISCUSSION PAPER SERIES*(11545).

McNamee, P., & Mendolia, S. (2018). Changes in health-related quality of life: a compensating income variation approach. *Applied Economics*, *00*(00), 1–12. doi: 10.1080/00036846.2018.1504160

Mogstad, M., Santos, A., & Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, *86*(5), 1589–1619. doi: 10.3982/ECTA15463

Neumann, P. J., Sanders, G. D., Russell, L. B., Siegel, J. E., & Ganiats, T. G. (2016). *Cost effectiveness in health and medicine*. New York: Oxford University Press.

Norman, R., Cronin, P., Viney, R., King, M., Street, D., & Ratcliffe, J. (2009). International Comparisons in Valuing EQ-5D Health States :. *Value in Health*, *12*(8), 1194–1200. doi: 10.1111/j.1524-4733.2009.00581.x

Ólafsdóttir, T., Ásgeirsdóttir, T. L., & Norton, E. C. (2020). Valuing pain using the subjective well-being method. *Economics and Human Biology*, *37*. doi: 10.1016/j.ehb.2019.100827

Oswald, A. J., & Powdthavee, N. (2008). Does happiness adapt? a longitudinal study of disability with implications for economists and judges. *Journal of Public Economics*, *92*(5-6), 1061–1077. doi: 10.1016/j.jpubeco.2008.01.002

Philipson, T. J., & Jena, A. B. (2006). Who benefits from new medical technologies? estimates of consumer and producer surpluses for hiv/aids drugs. In *Forum for health economics & policy* (Vol. 9).

Pinto-Prades, J.-L., Sánchez-Martínez, F.-I., Corbacho, B., & Baker, R. (2014). Valuing qalys at the end of life. *Social science & medicine*, *113*, 5–14. doi: 10.1016/j.socscimed.2014.04.039

Pischke, J.-S. (2011). Money and hapiness: Evidence from the industry wage structure. *NBER Working Paper*(17056).

Pischke, J.-S., & Schwandt, H. (2012). *A cautionary note on using industry affiliation to predict income* (Tech. Rep.). National Bureau of Economic Research.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection* (Vol. 1). Wiley Online Library.

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

Ryen, L., & Svensson, M. (2015). The willingness to pay for a quality adjusted life year: a review of the empirical literature. *Health Economics*, *24*(10), 1289–1301. doi: 10.1002/hec.3085

Sabatini, F. (2014). The relationship between happiness and health: evidence from italy. *Social Science & Medicine*, *114*, 178–187. doi: 10.1016/j.socscimed.2014.05.024

Socio-Economic Panel, S. (2019). *Socio-economic panel (soep), data for years 1984–2018,* *version 35.* SOEP.

Soeteman, L., van Exel, J., & Bobinac, A. (2017). The impact of the design of payment scales on the willingness to pay for health gains. *The European Journal of Health Economics*, *18*(6), 743–760. doi: 10.1007/s10198-016-0825-y

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

Vatter, J. (2012). Well-being in Germany: What explains the regional variation? *Discussion Papers, Forschungszentrum Generationenverträge der Albert-Ludwigs- Universität Freiburg*, *No. 50*.

Veenhoven, R. (2008, sep). Healthy happiness: effects of happiness on physical health and the consequences for preventive health care. *Journal of Happiness Studies*, *9*(3), 449–469. doi: 10.1007/s10902-006-9042-1

Verardi, V., & Croux, C. (2009, sep). Robust Regression in Stata. *The Stata Journal: Promoting communications on statistics and Stata*, *9*(3), 439–453. doi: 10.1177/1536867X0900900306

Wagstaff, A. (1991). Qalys and the equity-efficiency trade-off. *Journal of health economics*, *10*(1), 21–41. doi: 10.1016/0167-6296(91)90015-F

Ware, J. E., Keller, S. D., & Kosinski, M. (1995). *Sf-12: How to score the sf-12 physical and mental health summary scales.* Health Institute, New England Medical Center.

Woods, B., Revill, P., Sculpher, M., & Claxton, K. (2016). Country-Level Cost-Effectiveness Thresholds: Initial Estimates and the Need for Further Research. *Value in Health*, *19*(8), 929–935. doi: 10.1016/j.jval.2016.02.017

A1 Appendix - Dataset Details

Table A1: Dataset conditioning

Description	**Observations**
Total number of individual-year observations (2002-2018)	440,852
Clean Control variables: sex, age, disability, marital/employment status	326,717
Remove working individuals without industry-occupation information	309,253
At least two consecutively observed self-reported SF12	243,157
Remove observation not included in consecutive triplets	220,358
Excluding observations without lag of SF12, income and disability	186,906
Individual $*$ Years	186,906
Individuals	29,735
Without SF6D-Imputation	
Individual * Years	85,433
Individuals	21,718

A2 Appendix - Additional Results

Table A2: Subgroup results

	Baseline		Age<50		Age≥50		Male		Female	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Income in 1000's	0.05***	0.10***	0.07***	-0.01	0.03***	0.21***	0.04***	0.02	0.05***	0.20***
	(0.01)	(0.03)	(0.01)	(0.05)	(0.01)	(0.05)	(0.01)	(0.04)	(0.01)	(0.05)
Income in 1000's (t-1)	0.01	0.04	0.02**	0.14***	0.00	-0.06	0.01*	0.06*	0.00	0.02
	(0.01)	(0.03)	(0.01)	(0.04)	(0.01)	(0.04)	(0.01)	(0.04)	(0.01)	(0.04)
SF-6D utility	3.12***	3.12***	2.99***	3.00***	3.17***	3.16***	3.01***	3.01***	3.21***	3.21***
	(0.06)	(0.05)	(0.10)	(0.09)	(0.08)	(0.08)	(0.09)	(0.09)	(0.09)	(0.08)
SF-6D utility (t-1)	0.10*	0.10*	0.04	0.03	0.09	0.10	0.05	0.05	0.16*	0.14*
	(0.06)	(0.05)	(0.09)	(0.09)	(0.08)	(0.08)	(0.09)	(0.08)	(0.08)	(0.08)
Model statistics										
Cragg-Donald		1,863.7		661.1		265.5		513.5		398.7
Anderson		3,642.0		1,179.4		504.9		929.8		747.1
Endogeneity test		10.0		11.5		14.5		1.8		16.0
BIC	540,755	540,995	223,269	223,623	307,453	308,339	247,607	247,666	293,222	293,723
Observations	186,902	186,902	80,324	80,324	105,231	105,231	87,192	87,192	99,710	99,710
CIV in €	58,533	22,717	34,691	23,814	98,518	21,193	52,956	36,397	61,947	15,335

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. BIC Bayesian information criteria. East defined as former GDR.

Table A3: Results for unweighted and separate SF-6D levels

	Baseline				SF-6D sum score				SF-6D levels			
	OLS		IV		OLS		IV		OLS		IV	
Income in 1000's	0.05***	(0.01)	0.14***	(0.05)	0.05***	(0.01)	0.14***	(0.05)	0.05***	(0.01)	0.12**	(0.05)
Income in 1000's ($t-1$)	-0.00	(0.01)	-0.00	(0.07)	0.00	(0.01)	0.01	(0.07)	0.00	(0.00)	0.02	(0.07)
SF-6D utility	3.52***	(0.06)	3.51***	(0.05)								
SF-6D utility ($t-1$)	0.47***	(0.05)	0.46***	(0.05)								
SF-6D Summary Score					2.69***	(0.04)	2.69***	(0.04)				
SF-6D Summary Score ($t-1$)					0.30***	(0.04)	0.29***	(0.04)				
Physical Function 2									-0.05***	(0.01)	-0.06***	(0.01)
Physical Function 3									-0.19***	(0.02)	-0.19***	(0.02)
Role Function 2									-0.00	(0.01)	-0.00	(0.01)
Role Function 3									-0.18***	(0.02)	-0.18***	(0.02)
Role Function 4									-0.15***	(0.02)	-0.15***	(0.02)
Social Function 2									-0.07***	(0.01)	-0.07***	(0.01)
Social Function 3									-0.30***	(0.02)	-0.30***	(0.02)
Social Function 4									-0.63***	(0.03)	-0.63***	(0.03)
Social Function 5									-0.81***	(0.08)	-0.81***	(0.05)
Pain 2									-0.02	(0.01)	-0.02	(0.01)
Pain 3									-0.08***	(0.02)	-0.08***	(0.02)
Pain 4									-0.19***	(0.02)	-0.19***	(0.02)
Pain 5									-0.32***	(0.05)	-0.31***	(0.04)
Mental Health 2									-0.13***	(0.01)	-0.13***	(0.01)
Mental Health 3									-0.40***	(0.02)	-0.40***	(0.02)
Mental Health 4									-0.94***	(0.03)	-0.94***	(0.02)
Mental Health 5									-1.82***	(0.09)	-1.82***	(0.05)
Vitality 2									-0.04	(0.03)	-0.04	(0.03)
Vitality 3									-0.20***	(0.03)	-0.20***	(0.03)
Vitality 4									-0.43***	(0.03)	-0.43***	(0.03)
Vitality 5									-0.67***	(0.04)	-0.66***	(0.04)
Model statistics												
Cragg-Donald			192.1				153.8				192.2	
Anderson			382.2				299.0				382.4	
Endogeneity test			5.8				5.5				4.9	
BIC	236,338		236,538		234,751		234,933		230,943		231,104	
Observations	85,433		85,433		85,433		85,433		85,433		85,433	
CIV in €	80,522		28,130		58,083		20,266		79,037		27,869	
Physical Function									3,729		1,337	
Role Function									3,024		1,060	
Social Function									16,264		5,749	
Pain									6,316		2,218	
Mental Health									36,361		12,819	
Vitality									13,343		4,685	

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. BIC Bayesian information criteria. Baseline specification without imputation of SF-6D utilities. Sum score with range from 0 to 1.

A3 Health State Dependence - Detailed Information

We explored the potential relevance of the health state dependence of consumption utility on the estimation of CIV_{QALY} values by constructing a sub-sample of individuals transitioning between good and bad health ("health change sample"). To do so, we used the mental and physical SF-12 component scores as a universal health-state measure. The mental and physical component scores (MCS and PCS) range from 0 (worst) to 100 (best) and are normalized to have a mean of 50 and a standard deviation of 10 (Ware et al., 1995). In a first step, we calculated for each individual i their respective maximum and minimum reported mental ($MCS_i^{min/max}$) and physical ($PCS_i^{min/max}$) component score across periods. Individuals were included in the sample if they experienced an overall score difference of at least 10 ($MCS_i^{max} - MCS_i^{min} \geq 10$ and/or $PCS_i^{max} - PCS_i^{min} \geq 10$). Subsequently, we calculated the mean score ($\overline{MCS_i}$ and $\overline{PCS_i}$) for each individual. A health state for a given individual in period t was considered good if both mental and physical scores were greater or equal to the mean ($MCS_{it} \geq \overline{MCS}_i$ and $PCS_{it} \geq \overline{PCS}_i$), and bad if both were below ($MCS_{it} < \overline{MCS}_{it}$ and $PCS_i < \overline{PCS}_i$). Lastly we conditioned on the consecutive observation of health states with at least two periods spend in each to allow for a fixed effects estimation for good and bad health states separately. This reduced our sample considerably to 5,112 individuals. Table A4 compares the summary statistics for the main analysis sample and the health change sample.

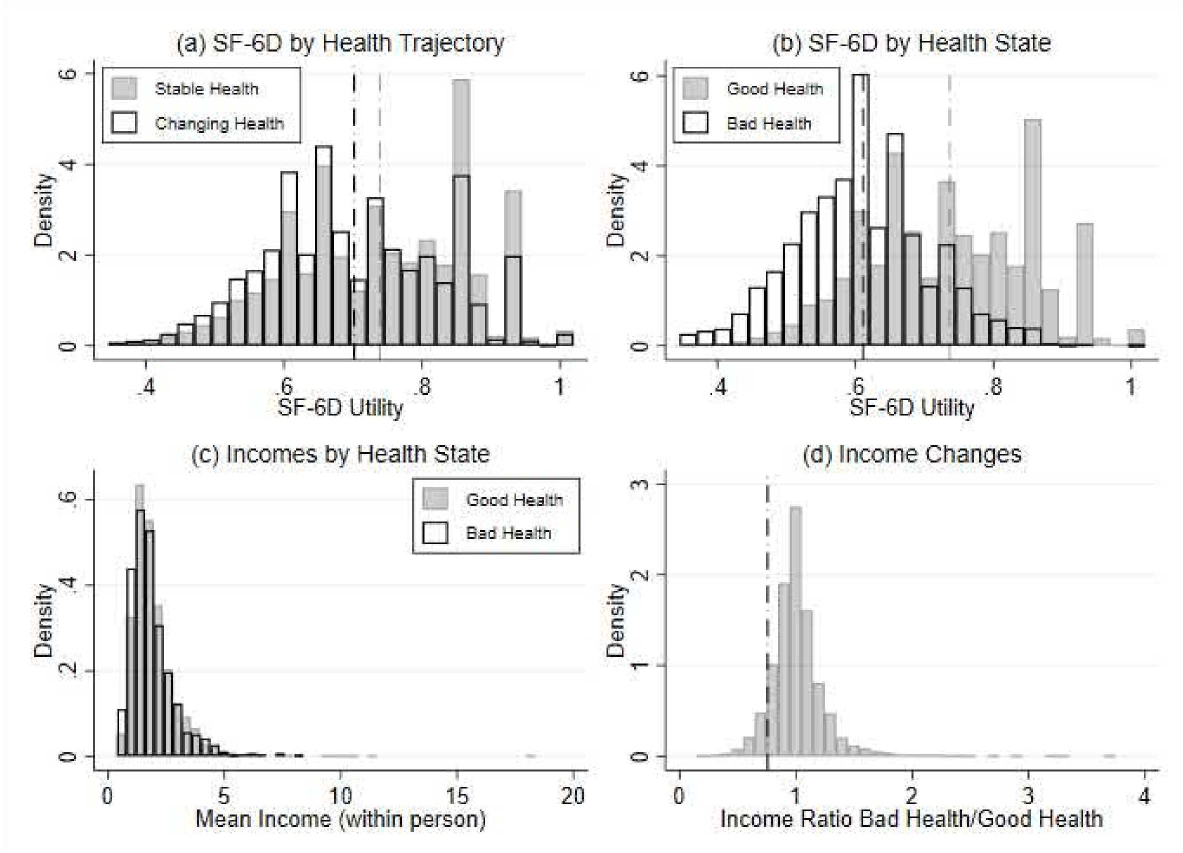
Table A4: Descriptive statistics - health state dependence sample

	All Individuals		Health Change Sample	
Variable	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
Life satisfaction	7.09	1.71	6.87	1.75
Income in 1000's	2.03	1.29	1.97	1.12
SF-6D utility	0.73	0.13	0.70	0.13
Disability	0.14	0.35	0.17	0.37
Age in years	53.67	15.78	56.33	15.51
(de facto) Married	0.67	0.47	0.68	0.47
Education: Primary	0.12	0.32	0.12	0.32
Education: Tertiary	0.63	0.48	0.65	0.48
Education: Secondary	0.25	0.43	0.23	0.42
Leisure time	2.18	2.03	2.35	2.10
Employed	0.56	0.50	0.50	0.50
Unemployed	0.04	0.21	0.04	0.21
Work hours	21.22	20.99	18.91	20.83
Tenure	7.03	9.96	6.59	9.95
Individuals * Years		186,902		48,861
Individuals		29,735		5,112

Overall there was a good overlap between the sample characteristics of the full analysis sample and the health change sample. The notable exceptions were slightly lower levels of life-satisfaction, income and SF-6D utility values and a slightly higher average age and disability-rate within the health change sample. These slight differences in age- and health-related variables were not surprising as we conditioned on individuals experiencing a substantial health change.

Figure A1 provides an overiew on the two most important variables for the CIV_{QALY} estimation; health and income. Panel (a) depicts the distribution of SF-6D utilities within the health change sample (black) and the rest of the sample (grey). Panel (b) depicts health utilities across health states within the health change sample with good health in grey and bad health in black. It becomes clear that our definition of bad health (based on MCS and PCS) coincides with substantially lower health utilities with the mean falling from 0.74 in good health to 0.61 in bad health.

Figure A1: Health change sample overview



Note: Panel (a) depicts distribution of SF-6D utility values and their means (dash-dotted lines) for individuals without a 10-point component score change over the observation period and individuals in the health-state dependence sample. Panel (b) depicts the distribution of SF-6D utilities of people in the health state dependence sample in their respective good and bad health states. Panel (c) plots the distribution of within-person mean equivalized household incomes in good and bad health states. Panel (d) depicts the ratio of equivalized housholed income in bad and good health states with the dash-dotted line marking the lower one standard-deviation.

As mentioned in the main body of the text, one concern when studying health state dependence of consumption utility is that there is the possibility that lower health is associated with a decrease in income leading to larger income coefficients. Figure A1 Panel (c) illustrates that this concern is warranted by plotting the within-person mean incomes across health states for the working population showing a shift of the distribution towards lower incomes.

Finkelstein et al. (2013) and Kools and Knoef (2019) addressed this directly using a two-step procedure to obtain income coefficients for individuals with stable incomes across health states. Our fixed-effects based approach does not allow for this. Instead we calculated the ratio between within-person income means in good and bad health states and excluded those individuals whose income-ratio was less than one standard deviation below the mean income-ratio. Panel

(d) plots these ratios and the corresponding cut-off point. This removes approximately 10% of working individuals from the sample. Table A5 presents the estimation results when excluding individuals with high income differences, which left a total of 4,656 observations.

Table A5: Health state dependence - excluding high income losses

	Baseline		**Good Health**		**Bad Health**	
	OLS	IV	OLS	IV	OLS	IV
Income in 1000's	0.07***	0.16**	0.06***	0.12	0.06	0.12
	(0.02)	(0.08)	(0.02)	(0.09)	(0.04)	(0.25)
Income in 1000's $(t-1)$	0.01	-0.01	0.02	0.00	0.03	0.06
	(0.01)	(0.07)	(0.02)	(0.07)	(0.03)	(0.21)
SF-6D utility	3.54***	3.54***	2.44***	2.43***	4.11***	4.09***
	(0.11)	(0.10)	(0.14)	(0.13)	(0.39)	(0.39)
SF-6D utility $(t-1)$	0.07	0.08	0.16	0.16	0.34	0.34
	(0.11)	(0.10)	(0.13)	(0.12)	(0.27)	(0.27)
Model statistics						
Cragg-Donald		575.5		415.9		81.6
Anderson		1,120.0		808.8		160.6
Endogeneity test		1.7		0.6		0.1
BIC	137,041	137,087	93,643	93,662	34,344	34,349
Observations	44,667	44,667	32,389	32,389	12,278	12,278
CIV in €	50,571	25,346	32,347	20,626	51,034	24,462

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

In a second step we further restricted the sample to only those individuals who worked during their respective participation in the panel. This was motivated by the fact that individuals might have moved into early retirement or unemployment following health changes, thereby decreasing the within-person income variance across periods. Table A6 depicts the estimation results after restricting the sample to the 3,032 individuals without large negative income changes and working throughout the entire observation period.

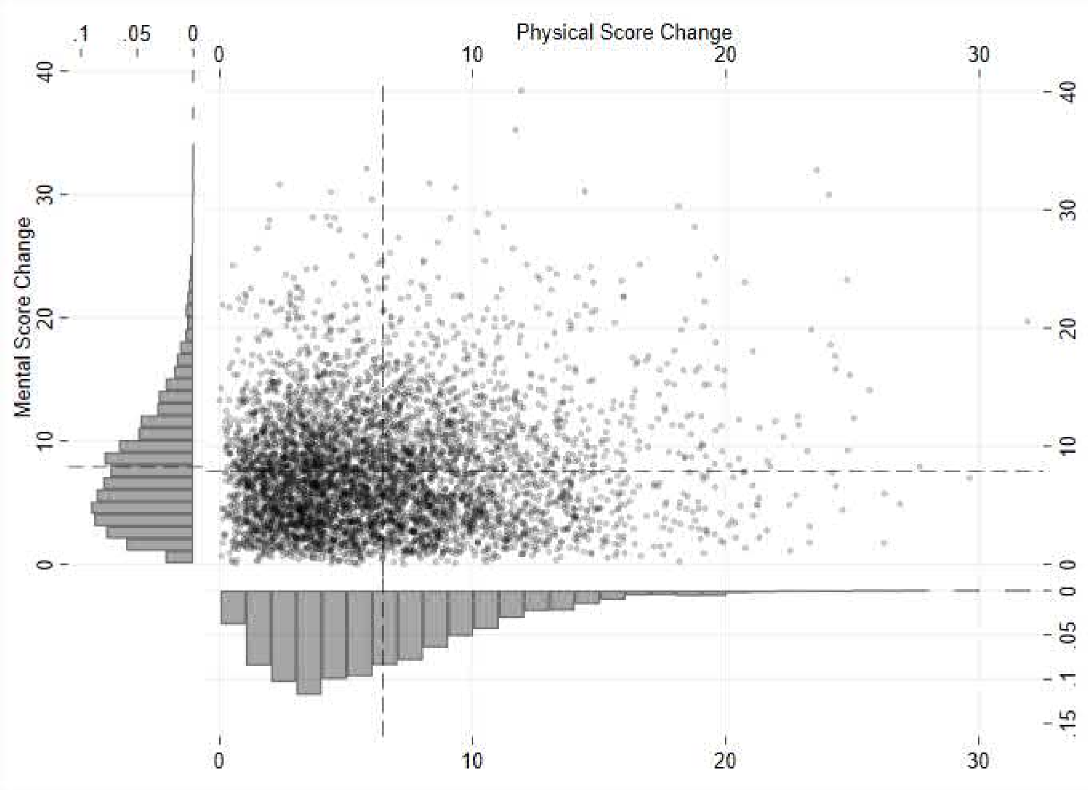
Table A6: Health State Dependence - excluding high income losses & unemployed/retired

	Baseline		**Good Health**		**Bad Health**	
	OLS	IV	OLS	IV	OLS	IV
Income in 1000's	0.07***	0.11	0.06***	0.04	0.08*	0.21
	(0.02)	(0.08)	(0.02)	(0.09)	(0.04)	(0.24)
Income in 1000's $(t-1)$	0.00	0.02	0.02	0.04	0.02	0.04
	(0.02)	(0.06)	(0.02)	(0.07)	(0.03)	(0.21)
SF-6D utility	3.45***	3.45***	2.51***	2.51***	3.52***	3.48***
	(0.13)	(0.12)	(0.17)	(0.15)	(0.46)	(0.46)
SF-6D utility $(t-1)$	0.08	0.09	0.23	0.23	0.41	0.42
	(0.13)	(0.12)	(0.15)	(0.14)	(0.32)	(0.33)
Model statistics						
Cragg-Donald		454.9		331.8		65.4
Anderson		880.9		642.1		128.4
Endogeneity test		0.5		0.1		0.3
BIC	89,392	89,401	61,756	61,758	21,876	21,892
Observations	29,508	29,508	21,530	21,530	7,978	7,978
CIV in €	46,735	28,776	37,299	36,176	41,513	15,599

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

In a last step we restrict our sample to individuals experiencing more severe and sudden health changes. Given our selection on a 10-point decrease in mental or physical component scores the sample might include individuals who experienced a gradual decrease in their health, which is substantial in absolute terms, but occurred over a long time-period. In such cases, the definition of health states based on the overall within-person mean wrongfully identifies individuals as switching between different health states despite the actual differences between both being rather small. Figure A2 plots the differences in mean mental ($\overline{MCS}_i^{good} - \overline{MCS}_i^{bad}$) and physical ($\overline{PCS}_i^{good} - \overline{PCS}_i^{bad}$) component scores and the raw distribution of differences for each health dimension for all 5,112 individuals. By construction higher score changes indicate worse health changes. The mean difference in mental health scores was 7.90 and for physical health 6.46, however there was a substantial number of individuals who experienced small health differences between both states.

Figure A2: Health change sample - mental/phsyical component score differences



Note: The main panel plots the individual-level differences between mean mental (vertical axis) and physical (horizontal axis) component score changes between good and bad health states. The raw distributions are plotted in the vertical and horizontal side panels with black dashed lines indicating mean score changes.

We explored to what extend the fact that some individuals in the health change sample only experienced a small or gradual health change across defined health states by excluding individuals for which the differences in mean component scores in both dimension was below 5 or one half standard deviation. Table A7 depicts the results when also excluding these individuals next to those with high income losses as well as consistently out of work, leaving only 2,567 individuals in the analysis sample. As expected the point estimates for the IV specification become very imprecise. Nonetheless, the general pattern indicating higher income coefficients in bad health states remains across specifications, suggesting the presence of positive health state dependency.

Table A7: Health State Dependence - Excluding high income losses, unemployed/retired & only severe health changes

	Baseline		Good Health		Bad Health	
	OLS	IV	OLS	IV	OLS	IV
Income in 1000's	0.07***	0.13	0.05**	0.06	0.08*	0.24
	(0.02)	(0.09)	(0.02)	(0.10)	(0.05)	(0.28)
Income in 1000's $(t-1)$	-0.01	0.01	0.00	0.02	-0.00	0.11
	(0.02)	(0.07)	(0.02)	(0.08)	(0.03)	(0.21)
SF-6D utility	3.64***	3.64***	2.64***	2.64***	3.65***	3.62***
	(0.14)	(0.13)	(0.19)	(0.17)	(0.49)	(0.50)
SF-6D utility $(t-1)$	0.05	0.06	0.22	0.22	0.50	0.50
	(0.14)	(0.13)	(0.17)	(0.16)	(0.34)	(0.35)
Model statistics						
Cragg-Donald		346.5		243.6		55.3
Anderson		672.9		473.2		108.9
Endogeneity test		1.1		0.1		0.5
BIC	74,566	74,586	50,415	50,416	19,184	19,225
Observations	24,272	24,272	17,414	17,414	6,858	6,858
CIV in €	68,898	27,828	54,494	35,519	54,631	11,890

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