

Following the Money: An Econometric Investigation into Health Valuations

Tinna Laufey Ásgeirsdóttir^{*1} and Viktor Andri Kárasón^{†1}

¹Department of Economics, University of Iceland, 101 Reykjavík, Iceland

July 2025

Work in progress

Abstract

In 2018, two studies valuing health-related quality of life using the same data (HILDA) and methods (Compensating Income Variation - CIV), produced irreconcilable results. One paper estimated a full QALY at A\$42,000–A\$67,000, while the other valued minor QALY changes as high as A\$162,000. We embark on econometric detective work to identify the sources of these discrepancies, carefully aligning samples, cross-examining methodologies, and uncovering methodological differences driving the divergent results. The results show that different income treatments drive these differences, highlighting the sensitivity of CIV results and the need for methodological consistency in this literature.

Keywords: Compensating Income Variation, Health, Income, QALY, Econometrics

JEL Codes: I31, D61, C83, I10

^{*}Corresponding author. Email: ta@hi.is, Tel: +354 865 0821.

[†]Email: vak17@hi.is, Tel: +354 858 7766.

1 Introduction

In recent years, a broad cross-disciplinary movement has sought to strengthen the *rigor, transparency, and replicability* of scientific research. Large multi-lab efforts in psychology revealed that fewer than half of prominent findings could be reproduced ([Open Science Collaboration, 2015](#)), while meta-research in biomedicine famously argued that many published claims are likely false-positive artifacts of flexible analyses ([Ioannidis, 2005](#)). Political science adopted DA-RT transparency standards to guard against similar risks ([Lupia and Elman, 2014](#)), and flagship reform papers in *Science* and *Nature Human Behavior* now refer to an emerging “credibility revolution” in science at large ([Nosek et al., 2015](#); [Munafò et al., 2017](#)). Across disciplines, journals increasingly mandate data-and-code deposits, preregistration platforms are commonplace, and dedicated replication consortia continue to grow.

Economics has joined this movement. For example, the causal-identification “credibility revolution” celebrated research designs that deliver persuasive causal inference ([Angrist and Pischke, 2010](#)). Similarly, a newer but fast-growing transparency push has reshaped publication norms: leading journals require public replication packages, the American Economic Association employs a Data Editor, and large-scale reproducibility audits find mixed success rates ([Chang and Li, 2018](#); [Christensen et al., 2019](#); [Institute for Replication, 2024](#)). These developments create demand for what we call *methodological forensics*: systematic scrutiny of identification and specification choices that can materially alter empirical conclusions. We contribute directly to this agenda.

An opportunity for methodological forensics arises from two papers published online within eight weeks of each other, yet were written in isolation. On *18 June 2018* [Huang et al. \(2018\)](#) published what they described as “*the first study that uses the wellbeing-valuation method to monetize the value of a QALY*,” while on *13 August 2018* [McNamee and Mendolia \(2018\)](#) claimed to be publishing “*the first set of monetary values for health losses using SF-6D utility values*.” Both studies draw on the same data—the Household, Income and Labour Dynamics in Australia (HILDA) panel—and both use the SF-6D index to translate health

states into QALY units. Their shared objective is identical: regress life satisfaction on health and income, then compute the compensating-income variation (CIV) that equates a given health change to its monetary counterpart.

One might therefore expect broadly similar valuations. Instead, the headline numbers appear irreconcilable. [Huang et al. \(2018\)](#) estimate A\$42,000-A\$67,000 (A\$52,547 and A\$83,357 at 2023 prices respectively) for a QALY gain, whereas [McNamee and Mendolia \(2018\)](#) report A\$39,486 (A\$55,107 at 2023 prices) for a 0.04 QALY loss and a staggering A\$162,211 (A\$226,384 at 2023 prices) for a 0.10 loss. Put differently, a one-tenth drop in QALY is valued at more than *three times* a full QALY gain. Both articles emphasize policy relevance, yet policymakers relying on one study versus the other would reach vastly different conclusions. Motivated by this discrepancy, we embark on an econometric “detective story”: we replicate each paper, place them on a common sample footing, and then swap their key specification choices to pinpoint the source of divergence.

Like any applied study, both papers necessarily make dozens of modeling choices, choices that, taken individually, look perfectly defensible. Researchers must decide which survey waves to include, how to code health scores, whether to equalize income, which shocks to treat as exogenous, how to specify the estimation equation, and which covariates or fixed effects to include. None of these decisions are obviously right or wrong *ex ante*; they are part and parcel of empirical practice.

Against that backdrop, the papers diverge along four main dimensions. **(i) Health metric.** [Huang et al. \(2018\)](#) value a *full-year* QALY gain using the raw SF-6D index, whereas [McNamee and Mendolia \(2018\)](#) value *marginal* SF-6D *losses* of varying sizes (-0.01 to -0.10) from different starting points. **(ii) Income treatment.** Both authors cite [Frijters et al. \(2011\)](#) to motivate an “exogenous shock” strategy, yet implement it differently: [Huang et al. \(2018\)](#) instrument *equivalized* household income with a *financial worsening* indicator in a two-stage least-squares (2SLS) design, whereas [McNamee and Mendolia \(2018\)](#) enter a *financial improvement* indicator directly in a reduced-form life-satisfaction equation and

later scale by average windfall income. **(iii) Time window.** Huang et al. (2018) estimate a two-year rolling panel and a “long-run equivalence” variant; McNamee and Mendolia (2018) use *contemporaneous* (one-year) changes. **(iv) Additional choices.** Wave coverage differs (2002–2015 vs. 2001–2010), as do certain controls: Huang et al. (2018) include year fixed effects, age, leisure capacity, and long-term health; McNamee and Mendolia (2018) add number of children, housing tenure, remoteness, and major negative life events. An overview of methodological similarities and differences across the two studies can be found in Table 1.

The orientation of the health and income shocks in Table 1 implies that Huang et al. (2018) estimate a *willingness to pay* (WTP) for a health gain, whereas McNamee and Mendolia (2018) estimate a *willingness to accept* (WTA) compensation for a health loss. A large literature shows that WTA valuations typically exceed WTP, with meta-analyses reporting median WTA/WTP ratios between about 1.5 and 5 (Hanemann, 1991; Horowitz and McConnell, 2002; O’Brien et al., 1998). While recognizing this pattern is important, even the upper end of the documented range cannot account for the forty-fold gap between the two studies examined here, so we treat orientation as a factor to bear in mind rather than the primary explanation for their divergence.

A systematic review by Ryen and Svensson (2015) illustrates just how wide the international literature already is: across 24 studies they find a trimmed *mean* willingness-to-pay (WTP) of €118,839 (2010 price level) per QALY, but a *median* of only €24,226 (2010 price level), evidence that extreme outliers coexist with more modest central tendencies. In that light, Huang’s estimate sits near the lower tail of published means, while McNamee–Mendolia’s marginal-loss numbers land in the upper outlier range. The question, therefore, is not whether either paper is “wrong” in some obvious sense, but *which modeling choice, or interaction of choices, drives the spread?* Given that the two studies being discussed here use the same data source and largely the same method. We aim to examine the potential reasons for these different results.

Our investigation proceeds in three stages. **First**, we replicate each study on the authors’

Table 1: Key methodological *similarities* and *differences* between Huang et al. (2018) and McNamee and Mendolia (2018)

Panel A: Similarities		
Aspect	Description (applies to <i>both</i> studies)	
Aim	Capture the monetary value of health-related quality of life (HRQoL) in QALYs.	
Data	<i>Household, Income and Labour Dynamics in Australia</i> (HILDA) longitudinal survey.	
Health instrument	QALY measures derived from the preference-based SF-6D index.	
Income endogeneity	Addressed with unexpected income shocks, citing Frijters et al. (2011) .	
Evaluation method	Compensating-income variation (CIV).	
Panel B: Differences		
Aspect	Huang et al. (2018)	McNamee and Mendolia (2018)
Health metric used	Full QALY gain ($\Delta = +1$)	Marginal QALY losses ($\Delta = -0.01$ to -0.10) from different initial health states
Income treatment	a) Dummy for <i>financial worsening</i> in last 12 months b) Instrumental variable estimation c) Income is equivalized	a) Dummy for <i>financial improvement</i> in last 12 months b) Reduced-form estimation c) Income is <i>not</i> equivalized
Estimation window	Two-year rolling window and long-run equivalence	Contemporaneous
Data waves	2002–2015 (waves 2–15)	2001–2010 (waves 1–10)

Note: Both studies condition on a broadly similar covariate set—marital status, education, employment status, and *individual* fixed effects. Huang et al. (2018) add age, leisure-capacity and a long-term-health indicator and include *year* fixed effects, whereas McNamee and Mendolia (2018) add the number of children, housing-tenure, geographic remoteness, and a small set of negative life-event dummies. These modest differences do not materially alter the common analytical core.

Because Huang model health *gains* and instrumented income with negative shocks, their CIV estimates correspond to a willingness-to-pay (WTP) interpretation; McNamee & Mendolia model health *losses* and a positive income shock, yielding a willingness-to-accept (WTA) interpretation.

own samples to rule out transcription or coding errors; in both studies, the published results are reproduced. **Second**, we re-estimate both models on an identical HILDA sample (waves 2–23), ensuring that discrepancies are methodological rather than data-driven. **Third**, based on an examination of the health and income coefficients explained below, we implement a “specification swap”: holding every other modeling choice constant, we replace the income treatment used by [Huang et al. \(2018\)](#) with that of [McNamee and Mendolia \(2018\)](#) and vice versa. These forensics exercises isolate the net contribution of the income specification.

The results reveal a clear pattern. Across all permutations, the estimated effect of health on life satisfaction is remarkably stable, but the income coefficient—and thus the CIV valuation—varies substantially with the sign and role of the income shock. When the income treatment is harmonized, the original valuation gap all but disappears; remaining differences arising from controls, time windows, or sample waves are small and partly offsetting. In short, simple econometric decisions explain almost the entire divergence.

These findings illustrate how otherwise routine decisions about methods related to income (instrumented versus reduced form, worsening versus improving income, and the use or omission of income equivalization) can exert a dominant influence on headline valuations, a cautionary tale for the growing CIV literature. Policymakers consulting one study or the other would therefore draw starkly different conclusions about cost-effectiveness thresholds or compensation formulas. By tracing the divergence in valuations to its underlying source, our methodological forensics contribute to the wider effort to make applied economic research more transparent, reliable, and policy-relevant.

2 Data

The HILDA survey is a comprehensive social and economic longitudinal survey, with a particular focus on family and household information, income, and work ([Watson and Wooden, 2012](#); [Summerfield et al., 2024](#)). We use waves 2–23 of the HILDA survey, corresponding to

the years 2002–2023.

Life satisfaction is used as our dependent variable. The respondents answer the question, “*All things considered, how satisfied are you with your life?*” on a scale from 0 to 10, where zero means “*totally dissatisfied*” and ten means “*totally satisfied*.” The distribution is negatively skewed, with 70% of responses being eight or above. The full distribution can be found in Appendix Figure A1.

The independent variables of interest are the SF-6D score (SF-6D health-utility values), dummy variables for different changes in the SF-6D score, household income, and a dummy variables for financial changes. The SF-6D score is derived from the SF-36 (Short Form-36) questionnaire, a standardized list of 36 questions originally constructed to survey health status in the Medical Outcomes Study (Ware and Sherbourne, 1992; Tarlov, 1989). The score comprises six different health dimensions: physical functioning, role limitations, social functioning, pain, mental health, and vitality. Together, scores from these six dimensions are combined into a 0–1 scale, where one represents full health and zero is equivalent to death. For the score to provide an accurate representation of each health state, a standard gamble method was used to rank health states against each other, resulting in different weights for each health state and possible scores ranging between 0.301 and 1 (Brazier et al., 2002). The distribution of SF-6D scores can be found in Appendix Figure A2.

Household income is measured in two different ways in the studies. McNamee and Mendolia (2018) use a dummy variable for a major financial improvement event as their income variable, while Huang et al. (2018) use equivalized household income instrumented by a major financial worsening event as their income variable. Therefore, we are interested in both financial improvement and worsening. The questions regarding these financial events are found under the major life events section, where respondents are asked, “*We would now like you to think about major events that have happened in your life over the past 12 months,*” and for each statement, they are tasked with checking a box for either “*yes*” or “*no*.” For the financial improvement, respondents are asked whether they have experienced “*major*

improvement in financial situation (e.g., won lottery, received an inheritance),” and for the financial worsening, they are asked if they have experienced *“major worsening in financial situation (e.g., went bankrupt).”* Out of 233,406 observations, there are 8,300 instances of financial improvements and 7,310 of financial worsening.

For equivalized household income, the OECD scale is used. The scale assigns a weight of 1 to the first adult in the household, 0.5 to every additional adult (aged 15 and over), and 0.3 to every child in the household. The age of individuals in the sample is based on their birth year and not their date of birth, as access to that information is restricted. Therefore, individuals born in January and December of the year 2000 are both assigned the age of 15 in the year 2015 and, therefore, classified as adults that year and onwards. The distribution of windfall income can be found in Appendix Figure A3, and the distribution of equivalized windfall income can be found in Appendix Figure A4.

We also use the same total set of control variables as each paper being replicated. The ones used in both papers are marital status, education, and employment status. In addition, we include age, leisure capacity, and long-term health conditions as controls, as Huang et al. (2018) do, and the number of children in the household, housing tenure, remoteness, and life events, as McNamee and Mendolia (2018) do. All observations with missing values were removed from our sample. This leaves us with 233,406 observations on 25,652 individuals out of 419,641 observations on 38,324 individuals in the HILDA dataset. Summary statistics can be found in Table 2.

3 Methods

In its simplest form, the standard approach for CIV calculations is to assume that life satisfaction (LS) is a function of the desiderata in question, here health (H), income (Y), and some other factors (X), so that:

$$LS = f(H, Y, X) \tag{1}$$

Table 2: Summary statistics, waves 2-23

Variable	Mean	SD	Measure
Life satisfaction	8.0	1.4	0 – 10, 0 totally dissatisfied, 10 totally satisfied
Household income (A\$1000) ^a	154.823	152.784	–3682.045 – 2889.373
Equivalized household income (A\$1000) ^a	84.735	81.567	–3064.276 – 2889.373
Household windfall income (A\$1000) ^{ab}	165.402	295.999	0.011 – 2844.958
Household lost income (A\$1000) ^{ac}	–63.113	126.522	–2340.867 – –0.008
SF-6D	0.757	0.123	0.301 - 1
Long-term condition	0.281	0.449	0 no, 1 yes
Financial worsening	0.026	0.158	0 no, 1 yes
Financial improvement	0.030	0.171	0 no, 1 yes
Age ^d	46.6	18.1	16 - 102
Married/de facto	0.659	0.474	0 no, 1 yes
Education	1.0	0.7	0 – 2, 0 less than 12 years, 1 12 years or equivalent, 2 12 years or greater
Leisure capacity ^e	0.336	0.452	0 – 1, 0 no time spent away from paid employment, 1 no time spent in paid employment
Unemployment	0.030	0.171	0 no, 1 yes

^aAll incomes are in Australian dollars, converted to the 2023 price level.

^bIncrease in gross household income between years t and $t-1$ for those who report financial improvement in the previous 12 months in year t .

^cDecrease in gross household income between years t and $t-1$ for those who report financial worsening in the previous 12 months in year t .

^dAge is reported in year t for observations in years t and $t-1$.

^eLeisure capacity is approximated using the inverse of "Percent time spent in jobs last financial year".

which translates to the following estimation equation:

$$LS = \alpha + \beta_1 H + \beta_2 Y + \beta_3 X + \epsilon \quad (2)$$

When using a continuous income variable, the trade-off value is calculated as follows:

$$\text{CIV} = -\frac{\beta_1}{\beta_2} \quad (3)$$

Equations (1)–(3) provide the single-period benchmark for CIV. Both studies examined here begin from this template but then extend it differently. [Huang et al. \(2018\)](#) embed a two-year rolling window with continuous income and a long-run equivalence formulation, while [McNamee and Mendolia \(2018\)](#) rely on a contemporaneous income-shock dummy. Those modeling choices alter the algebra and yield distinct CIV expressions as explained in further detail below, and in Appendix A.2, which derives them step by step.

We undertake a three-stage investigation to uncover what drives the seemingly irreconcilable results across the two studies. First, we do basic replications using the methods described in each paper and the waves of data used by the authors. This should establish whether the different results stem from a mistake in either paper. Second, we replicate the studies using the original methods from each paper, but with a fuller set of data and the same sample across both studies, that is, from waves 2–23 of the HILDA survey ([Watson and Wooden, 2012](#); [Summerfield et al., 2024](#)). Given that the differences across results remain, we examine what we deem the highest likelihood cause of the discrepancies, based on a thorough examination of the estimated coefficients.

Specifically, we asked which part of the two models *could* plausibly drive the valuation gap using a “coefficient sanity check”. This points to income rather than health for the following reason. For health, [Huang et al. \(2018\)](#) estimate a coefficient of 2.26 for a full-year QALY gain, whereas [McNamee and Mendolia \(2018\)](#) obtain 0.192 for a $\Delta = -0.10$ SF-6D loss. The [McNamee and Mendolia \(2018\)](#) coefficient is thus roughly 8.5% of the [Huang et al.](#)

(2018) coefficient and therefore close to the mechanical one-tenth scaling we would expect if both papers were pricing the same latent health construct.

In contrast, the income coefficients are orders of magnitude apart. Huang et al. (2018) report $\beta_Y = 0.080$ for a A\$1,000 (equivalized) change in annual income, while the windfall coefficient from McNamee and Mendolia (2018) is $\beta_{\text{windfall}} = 0.126$ for a one-off gain averaging A\$106,718. Expressed per A\$1,000, the latter equals only 0.0012, which is less than two per cent of the Huang et al. (2018) estimate. This stark mismatch makes income the prime suspect, as opposed to health. Therefore, we keep the health specification of each paper intact and perform an “income swap”: The models from Huang et al. (2018) are re-estimated using the reduced-form estimation approach, with the windfall dummy as in McNamee and Mendolia (2018), and vice versa. In other words, we thus replicate both papers as before, but using the income approach from Huang et al. (2018) when replicating McNamee and Mendolia (2018) and using the income approach from McNamee and Mendolia when replicating Huang et al. By comparing the results between the second and third replications, we can assess the impact of the income specification on each study. If valuation gaps remain after this swap, then other explanations would need to be examined.

3.1 Specifications for replicating Huang et al. (2018)

While replicating Huang et al. (2018), the calculations follow the steps listed in Equations 1, 2, and 3, although their model is based on a 2-year rolling window, individual and year fixed effects as well as an instrumental variable for household income. The estimation equation can thus be written as:

$$LS_{it} = \alpha + \beta_1 H_{it} + \beta_2 H_{it-1} + \beta_3 \hat{Y}_{it} + \beta_4 \hat{Y}_{it-1} + \beta_5 X_{it} + \lambda_i + \sigma_t + \epsilon_{it} \quad (4)$$

where H is the SF-6D score, \hat{Y} is equivalized household income, instrumented by financial worsening, X is a vector of individual control variables, and λ_i and σ_t are individual and year

fixed effects, respectively. A financial worsening event instruments income, FW_{it} , FW_{it-1} , $FW_{i,t-}$, where FW_{it} , denotes whether an event happened in the previous 12 months, $FW_{i,t-1}$ whether an event took place in the 12 months before, and $FW_{i,t-}$ whether an event took place in any year preceding the past two years.

The CIV values for health changes are calculated in the 2-year rolling window, where changes in the current year weigh twice as much as changes in the year prior:

$$CIV_{2\text{-year rolling window}} = \frac{2\beta_1 + \beta_2}{2\beta_3 + \beta_4} \quad (5)$$

For sustained changes in the long-run equivalence, each year is treated equally:

$$CIV_{\text{long-run equivalence}} = \frac{\beta_1 + \beta_2}{\beta_3 + \beta_4} \quad (6)$$

We include results for the 2-year rolling window and the long-run equivalence coefficients for completeness, although greater focus will be given to the 2-year rolling window in the interest of streamlining the discussion.

3.2 Specifications for replicating **McNamee and Mendolia (2018)**

The replications of **McNamee and Mendolia (2018)** also follow a similar process using Equations 1, 2, and 3 but differ slightly in the calculation of the CIV value. The estimations use individual fixed effects, and the estimation equations can be written as:

$$LS_{it} = \alpha + \beta_1 H_{it} + \beta_2 Y_{it} + \beta_3 X_{it} + \lambda_i + \epsilon_{it} \quad (7)$$

where H_{it} is a vector of health changes in different places on the SF-6D scale in one model and a vector of changes of different sizes in another, Y_{it} is an indicator variable for a financial improvement event, and X_{it} is a vector of control variables.

The health changes used in these models are dummy variables for different changes. In

the first model, there are dummy variables for specific changes in SF-6D, from 1 to 0.9; from 0.9 to 0.8; from 0.8 to 0.7; from 0.7 to 0.6; from over 0.9 to 0.8; from over 0.8 to 0.7; from over 0.7 to 0.6; from over 0.6 to a value less than 0.6. In the second model, the dummy variables are for negative changes of different sizes, regardless of the position on the scale, -0.01 ; -0.02 ; -0.03 continuing all the way to -0.1 , and then for changes greater than -0.1 . Full results are reported, although main focus will be on decreases in QALY by 0.1 in the interest of streamlining the discussion.

Since we are not using a continuous income variable but a dummy variable for financial improvements, the CIV calculations can be described using the following function:

$$CIV = -\frac{\beta_1}{\beta_2} \times \bar{Y}_{\text{wind}} \quad (8)$$

where \bar{Y}_{wind} is the average windfall income associated with a major improvement in finances. The average windfall income associated with this event is calculated as the average increase in gross household income of working-age individuals (18–70) in the estimation sample between years t and $t - 1$ where financial improvements are reported in year t .

3.3 Replications with switched income approaches

To preserve as much of the original methods as possible when income approaches are switched, we do the following. In [Huang et al. \(2018\)](#), the model is set up as a 2-year rolling window or a long-run equivalence with an income variable for each year. We keep the two-year rolling window and long-run equivalence in order to keep the theoretical specification stable and only alter the income treatment variables by now using dummy variable for major financial improvements each year in a reduced form empirical specification. To then calculate the monetary value of the 2-year rolling window and long-run equivalence health changes, we use the coefficients as before and multiply them by the average windfall income from the household, and keep this unequivalized in accordance with [McNamee and Mendolia \(2018\)](#).

The same goes when replicating [McNamee and Mendolia \(2018\)](#) with the income approach used by [Huang et al. \(2018\)](#). We want to preserve as much of the original method as possible while only changing the income treatment. To do that, we use financial worsening as an instrument for equivalized household income, but instead of using financial worsening in the same year, the year before, and any year prior to two years ago, we only use financial worsening in the same year and any year prior to that year, and thus keep the concurrent theoretical specification in [McNamee and Mendolia \(2018\)](#). The CIV is then estimated with the same coefficients as before, without multiplying it with average windfall income, since the income variable is now continuous and in monetary terms already.

Despite the different approaches, both studies reference [Frijters et al. \(2011\)](#) as justification for the way they approach their income variables. There, the effects on life satisfaction from financial worsening and improvement are studied using the HILDA data. Although that study suggests using financial improvement as the income variable and multiplying the dollar amount associated with that by the trade-off ratio, it also mentions that this method is similar to using a financial event as an instrument.

4 Results

The basic replications of [Huang et al. \(2018\)](#) and [McNamee and Mendolia \(2018\)](#) are reported in Appendix Tables [A1–A7](#). We follow the methods of each paper as closely as possible based on the information provided in the respective articles, using the same HILDA survey waves, variables, and methods as in the original publications. Our results are broadly consistent with those reported in the published papers. Naturally, some discrepancies arise because the original studies do not provide details on every decision made during data handling and analysis. However, this is to be expected. The results are similar enough that the differences between [Huang et al. \(2018\)](#) and [McNamee and Mendolia \(2018\)](#) do not appear to stem from mistakes in either paper. The results remain irreconcilable across papers based on these most

basic of replications. They do, however, use different samples in the original papers, and thus we also did in these base replications. The next step is thus to see if the differences in the samples and waves used could be the reason for the different results. In the following, we first present the additional replications for Huang et al. and subsequently the replications for McNamee and Mendolia, using the same sample for both studies. Each replication is compared to the original results in the corresponding paper, with price levels converted to 2023 prices.

4.1 Replications of Huang et al. (2018)

The replications of Huang et al. (2018) are reported in Tables 3 and 4. The first two columns present original results by Huang et al., with and without the instrumentation of income, respectively. Columns three and four show results from our replication, with and without financial worsening as an instrument using waves 2–23. Lastly, in the fifth column, we introduce the income-switching. That is, we run the model as in the replication, except with income treated as in McNamee and Mendolia (2018). That is, using financial improvement directly in the estimation equation instead of equivalized household income (instrumented or not). Relevant first-stage regressions can be found in Appendix Table A8.

4.2 Replications of McNamee and Mendolia (2018)

Replications of McNamee and Mendolia (2018) are presented in Tables 5 and 6. Table 5 shows the estimation results for changes in SF-6D scores to specific positions on the scale, while Table 6 shows the estimation results for changes of specific sizes in SF-6D scores. The first column of each table show the original results by McNamee and Mendolia. The subsequent column shows results from our replication using waves 2–23, and finally replications in which the income approach from Huang et al. (2018) is used. Although McNamee and Mendolia (2018) do not specify (a) whether or (b) to which year dollar amounts are converted, we assume that they use 2010 prices since their study is based on ten waves of HILDA data, the

Table 3: Estimated effects on life satisfaction, replication of Huang et al. (2018)

	Huang et al., no IV (2018) ^a	Huang et al., with IV (2018) ^a	Replication, no IV	Replication with IV	Replication using the RF method
SF-6D	2.432*** (0.037)	2.258*** (0.134)	2.470*** (0.027)	2.372*** (0.118)	2.472*** (0.027)
SF-6D a year ago	0.749*** (0.037)	0.778*** (0.136)	0.774*** (0.026)	0.758*** (0.119)	0.779*** (0.026)
Income in 1000's	0.0006*** (0.0001)	0.080*** (0.010)	0.0003*** (0.00004)	0.062*** (0.007)	
Income in 1000's a year ago	0.0004*** (0.0001)	-0.035*** (0.010)	0.0002*** (0.00004)	-0.038*** (0.008)	
Financial improvements					0.152*** (0.012)
Financial improvements a year ago					0.067*** (0.012)
Number of observations	134,224	106,457	233,406	233,406	233,406
R^2			0.061	0.001	0.061
Adjusted R^2			-0.055	-0.123	-0.055

Standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

^aSignificance at 10% level is not reported in [Huang et al. \(2018\)](#) and therefore not reported here.

Table 4: Monetary estimations from Huang et al. (2018) replication

	Huang et al., no IV (2018) ^a	Huang et al., with IV (2018)	Replication, no IV	Replication with IV	Reduced-form method w. mean income	Reduced-form method w. equivalized mean income
1 QALY 2-year rolling window	4,363,128	52,547	6,344,340	63,774	2,555,031	1,535,249
1 QALY long-run equivalence	3,956,276	83,357	5,818,731	128,837	2,456,307	1,475,929

All values have been converted to 2023 Australian dollars.

^aMonetary estimations for models without IV are not reported in the original paper by [Huang et al. \(2018\)](#). We have therefore calculated the monetary estimations ourselves to use in this comparison using the coefficients estimated by [Huang et al. \(2018\)](#).

final wave thus being 2010. It is however possible that no price adjustments were made, we do not expect this to substantially affect our results, given that the decade under examination in [McNamee and Mendolia \(2018\)](#) was not a high-inflation period. We converted all prices to the 2023 price level under that assumption for a better comparison with our results. Their mean windfall income is then A\$148,937, a similar amount to the mean windfall income in waves 2–23, which is A\$165,402.

When examining the monetary valuations of health changes to specific positions in Table 7, a big difference is apparent between each method. The biggest difference between the mean windfall replication and the replication using the IV method is by a factor of 66 when the SF-6D score goes from above 0.8 to 0.7 between years (A\$125,617 in our replication using mean windfall and A\$1,902 in our replication using IV). The smallest difference between the mean windfall replication and the replication using the IV method, where both results are significantly different from zero, is at the upper end of the SF-6D score, where going from 0.9

Table 5: Estimated effects of specific SF-6D QALY scores on life satisfaction, replication of McNamee and Mendolia (2019)

Changes in SF-6D QALY scores from t_{-1} to t	McNamee and Mendolia (2019)	Replication, RF	Replication, no IV	Replication with IV
No change	0.006 (0.009)	-0.065*** (0.005)	-0.066*** (0.005)	-0.078*** (0.016)
SF-6D from 0.7 to 0.6	-0.242** (0.025)	-0.259*** (0.013)	-0.264*** (0.013)	-0.302*** (0.039)
SF-6D from 0.8 to 0.7	-0.119** (0.016)	-0.126*** (0.009)	-0.127*** (0.009)	-0.137*** (0.029)
SF-6D from 0.9 to 0.8	-0.011 (0.012)	-0.032*** (0.007)	-0.031*** (0.007)	-0.073*** (0.026)
SF-6D from 1 to 0.9	-0.001 (0.032)	0.064*** (0.020)	0.065*** (0.020)	-0.003 (0.071)
SF-6D from > 0.7 to 0.6	-0.230** (0.022)	-0.259*** (0.012)	-0.261*** (0.012)	-0.307*** (0.033)
SF-6D from > 0.8 to 0.7	-0.108** (0.026)	-0.121*** (0.015)	-0.120*** (0.015)	-0.089** (0.044)
SF-6D from > 0.9 to 0.8	-0.002 (0.057)	0.005 (0.034)	0.009 (0.034)	-0.050 (0.111)
Other neg. changes to < 0.6	-0.590** (0.035)	-0.522*** (0.019)	-0.525*** (0.019)	-0.475*** (0.038)
Financial improvements	0.124** (0.022)	0.159*** (0.013)		
Income in 1000's			0.0003*** (0.00004)	0.047*** (0.006)
Number of observations	83,177	233,406	233,406	233,406
R^2		0.026	0.027	0.001
Adjusted R^2		-0.094	-0.093	-0.123

Robust standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

Table 6: Estimated effects of different size changes in SF-6D QALY scores on life satisfaction, replication of McNamee and Mendolia (2019)

Changes in SF-6D QALY scores from t_{-1} to t	McNamee and Mendolia (2019)	Replication, RF	Replication, no IV	Replication with IV
No change	−0.031** (0.010)	−0.016*** (0.006)	−0.016*** (0.006)	−0.042** (0.019)
Change equal to −0.01	−0.130** (0.030)	−0.157*** (0.019)	−0.157*** (0.019)	−0.076* (0.045)
Change equal to −0.02	−0.130** (0.028)	−0.090*** (0.014)	−0.092*** (0.014)	−0.140*** (0.048)
Change equal to −0.03	−0.151** (0.040)	−0.152*** (0.023)	−0.153*** (0.023)	−0.168*** (0.059)
Change equal to −0.04	−0.046** (0.012)	−0.024*** (0.007)	−0.024*** (0.007)	−0.042* (0.025)
Change equal to −0.05	−0.185** (0.033)	−0.140*** (0.019)	−0.142*** (0.019)	−0.166*** (0.051)
Change equal to −0.06	−0.147** (0.019)	−0.133*** (0.011)	−0.135*** (0.011)	−0.160*** (0.032)
Change equal to −0.07	−0.136** (0.035)	−0.142*** (0.019)	−0.143*** (0.019)	−0.116** (0.052)
Change equal to −0.08	−0.102** (0.023)	−0.109*** (0.013)	−0.110*** (0.013)	−0.117*** (0.039)
Change equal to −0.09	−0.207** (0.037)	−0.197*** (0.022)	−0.199*** (0.022)	−0.161*** (0.051)
Change equal to −0.1	−0.192** (0.030)	−0.141*** (0.017)	−0.142*** (0.017)	−0.162*** (0.056)
Change < −0.1	−0.228** (0.013)	−0.218*** (0.007)	−0.219*** (0.007)	−0.221*** (0.019)
Financial improvements	0.126** (0.022)	0.161*** (0.013)		
Income in 1000's			0.0003*** (0.00004)	0.048*** (0.006)
Number of observations	83,177	233,406	233,406	233,406
R^2		0.022	0.023	0.001
Adjusted R^2		−0.099	−0.098	−0.123

Robust standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

to 0.8 differs by a factor of 21 (A\$33,287 in our replication using mean windfall and A\$1,557 in our replication using IV).

Table 7: Monetary estimattions for SF-6D changes between different positions, replication of McNamee and Mendolia (2019)

Changes in SF-6D QALY scores from t_{-1} to t	McNamee and Mendolia (2019)	Reduced-form method mean	Reduced-form method equivalized mean	Replication, no IV	Replication with IV
SF-6D from 0.7 to 0.6	290,428	269,613	162,003	802,856	6,429
SF-6D from 0.8 to 0.7	142,979	131,032	78,734	387,993	2,917
SF-6D from 0.9 to 0.8	13,405	33,287	20,001	94,363	1,557
SF-6D from 1 to 0.9	1,192	−66,672	−40,061	−198,682	62
SF-6D from > 0.7 to 0.6	275,534	269,541	161,960	795,783	6,528
SF-6D from > 0.8 to 0.7	129,576	125,617	75,480	366,336	1,902
SF-6D from > 0.9 to 0.8	1,489	−4,991	−2,999	−27,985	1,071
Other neg. changes to < 0.6	707,453	542,852	326,185	1,597,970	10,109

All values have been converted to 2023 Australian dollars.

When examining the valuations of changes of different sizes in Table 8, the same pattern emerges. However, the difference between the replication using mean windfall and the replication using the IV method is slightly grater, with the biggest difference being by a factor of 102, for a change equal to -0.01 (A\$161,591 in our replication using mean windfall and A\$1,578 in our replication using IV) and the smallest difference being by a factor of 28, for a change equalling -0.04 (A\$24,906 in our replication using mean windfall and A\$880 in our replication using IV).

4.3 What matters most in the income treatment

Sections 4.1 and 4.2 have shown that, once health specification and sample are harmonized, the vast divergence between the two papers being examined is driven almost entirely by how income is modelled. Yet this “income treatment” comprises *three* separate modelling

Table 8: Monetary estimattions for SF-6D changes of different sizes, replication of McNamee and Mendolia (2019)

Changes in SF-6D QALY scores from t_{-1} to t	McNamee and Mendolia (2019)	Reduced-form method mean	Reduced-form method equivalized mean	Replication, no IV	Replication with IV
Change equal to -0.01	151,916	161,591	97,095	478,517	1,578
Change equal to -0.02	153,406	93,148	55,970	280,707	2,918
Change equal to -0.03	178,725	156,872	94,260	464,099	3,498
Change equal to -0.04	55,107	24,906	14,965	72,477	880
Change equal to -0.05	220,428	144,370	86,748	431,995	3,462
Change equal to -0.06	174,257	136,992	82,315	409,674	3,330
Change equal to -0.07	160,852	146,262	87,885	433,962	2,412
Change equal to -0.08	120,640	112,397	67,536	333,201	2,432
Change equal to -0.09	244,258	203,115	122,046	604,044	3,350
Change equal to -0.1	226,384	145,636	87,509	430,828	3,380
Change < -0.1	269,576	224,418	134,847	664,827	4,597

All values have been converted to 2023 Australian dollars.

choices that must be unpacked to see which one matters most. Table 9 packs the three modelling decisions related to income that differ between Huang et al. (2018) and McNamee and Mendolia (2018) into a single matrix: (a) whether the income change is a loss or a gain, (b) IV versus RF, and (c) equalisation of income (E) or not (U). The six columns form a 2×3 block for each time horizon (2-year rolling window vs. long-run equivalence). Reading *down* a column compares losses with gains; reading *across* columns first isolates the effect of switching from IV(E) to RF(E) (model only) and then the effect of moving from RF(E) to RF(U) (equalisation only). Derived ratios that show the comparative magnitudes of these margins are grouped underneath the raw valuations for quick reference.

Within every model–income definition, the *Imp./Wors.* ratio is by far the largest multiplier, ranging from **7.8** to **15.7**. Thus, the differences in the monetary valuations of health are first and foremost driven by whether the income change is perceived as a windfall or a setback. Holding the sign of the income change constant but switching from IV to RF (com-

paring IV(E) with RF(E)) boosts valuations by factors of about **1–3**. Comparing **RF(U)** with **RF(E)** shows that dropping the equivalisation roughly **doubles** the valuations with ratios of 1.66–1.85, i.e. equivalising income attenuates the RF estimates to about 54 % of their unequivalized level. Equivalisation, therefore, matters, but much less than the sign of the shock or the econometric strategy.

It is a coincidence in this case that each modelling choice made differently across studies pushes in the same direction: the [Huang et al. \(2018\)](#) configuration (IV, loss, E) always gives the *smallest* valuations, while the [McNamee and Mendolia \(2018\)](#) configuration (RF, gain, U) always gives the *largest*. The *cross ratios* in the last row make this explicit: the estimate by [McNamee and Mendolia \(2018\)](#) is up to **40 times** larger in the 2-year rolling-window model and **19 times** larger in the long-run equivalence model than estimates by [Huang et al. \(2018\)](#).

Table 9: Monetary valuation of a QALY under alternative income treatments

	2-year rolling window			Long-run equivalence		
	IV (E)	RF (E)	RF (U)	IV (E)	RF (E)	RF (U)
Financial worsening	63 774	166 861	309 057	128 837	162 190	300 406
Financial improvement	998 136	1 535 250	2 555 031	1 009 457	1 475 929	2 456 307
Imp./Wors. ratio [†]	15.65	9.20	8.27	7.84	9.10	8.18
RF(E)/IV(E) [‡] — worsening	—	2.62	—	—	1.26	—
RF(E)/IV(E) [‡] — improvement	—	1.54	—	—	1.46	—
RF(U)/RF(E) [§] — worsening	—	—	1.85	—	—	1.85
RF(U)/RF(E) [§] — improvement	—	—	1.66	—	—	1.66
Cross ratio [§]	—	—	40.06	—	—	19.07

Column keys: **IV (E)** = instrumental-variables model with *equivalized* income; **RF (E)** = reduced-form model with *equivalized* income; **RF (U)** = reduced-form model with *unequivalized* income. [†]Within-column ratio of valuations for income *improvements* to *worsenings* (proxy for WTA/WTP). [‡]Ratio RF(E) ÷ IV(E); reported separately for financial worsening and improvement. [§]Ratio RF(U) ÷ RF(E); reported separately for financial worsening and improvement. [§]Cross ratio = valuation under RF(U) with a financial *improvement* ÷ valuation under IV(E) with a financial *worsening* — combining all three modelling differences. All values are compensating-income variations (CIVs) in 2023 A\$. “—” indicates ratios that are undefined by construction.

5 Discussion

This study set out to investigate why two credible papers, both using the same Australian panel and broadly the same wellbeing-valuation framework, produced QALY valuations that differed by an order of magnitude. Our detective exercise reveals that, once health and sample choices are aligned, almost the entire gap can be traced to differences in how income is treated. This finding offers a cautionary tale for the broader CIV literature and, more generally, underscores how small, defensible modeling decisions can snowball into large, policy-relevant discrepancies, a central concern of the growing push for methodological forensics in economics.

Our evidence confirms that both studies are *reproducible*: rerunning the analyses described in each paper on the authors’ own wave selection, and on an expanded common sample, largely reproduces the published point estimates and the expected signs of health and income. Only when we *switch* the income specification, placing the windfall dummy from McNamee and Mendolia (2018) inside the model from Huang et al. (2018) and vice versa, does the gap collapse, while the SF-6D coefficients remain stable. Decomposing that effect further reveals that the *orientation* of the shock (financial improvement versus worsening) is the most significant factor, followed by the estimation method (RF vs IV), and whether income is adjusted for family size using the OECD modified equivalence scale. The implied WTA–WTP spread is far larger than typical ratios reported in the stated-preference literature, suggesting that CIV estimates may be especially sensitive to how income shocks are framed and operationalized.

Although it is impossible to do precisely, it is nonetheless instructive to gauge to what extent the income treatment explains the large difference in results between Huang et al. (2018) and McNamee and Mendolia (2018). Our replication of Huang et al. (2018) estimates the value of a QALY at A\$63,774. Now, it should be kept in mind that McNamee and Mendolia (2018) estimate the value of smaller changes in HRQoL, but summing the four estimated 0.1 changes in HRQoL reported in Table 5 yields a 0.4 change valued at

A\$367,260. This is much larger than the value of a full QALY under [Huang et al. \(2018\)](#), let alone if we were to compare only 40% of that number, or A\$25,510. The difference is a whopping A\$341,750 using the 2-year rolling window measure, or A\$315,725 using the long-run equivalence measure.

Switching to an IV method, using positive income shocks, and adjusting for household size more than accounts for this discrepancy. In fact, summing the value of the same four 0.1 HRQoL changes under the revised specification reduces the total valuation to just A\$10,965. This alternative treatment of income thus more than eliminates the difference across studies. While this thought experiment is not meant to be precise—due to other methodological differences—it nonetheless demonstrates that the bulk of the divergence is likely attributable to how income is incorporated into the estimation models.

This, in turn, points to a deeper explanation rooted in economic theory. One possible reason for the discrepancies could be the declining marginal utility of income or loss aversion. While one of the replicated papers emphasizes negative income shocks by instrumenting household income with financial worsening, the other focuses on positive income shocks by estimating the effects of financial improvements on life satisfaction. This asymmetry in treatment could result in an overestimation on one end and an underestimation on the other. It is important to note that both approaches to addressing income endogeneity have precedent in the literature and cannot be easily dismissed as inappropriate. One can be interpreted as estimating willingness to pay (WTP), while the other captures willingness to accept (WTA), and as demonstrated by [Hanemann \(1991\)](#), the difference between WTP and WTA can be substantial if no substitutes are available.

Another difference in approach to income endogeneity is that one set of authors use the income shock directly in the main reduced-form estimation equation, while the others incorporate it into a two-stage least squares instrumental-variables estimation. Both are valid strategies used by various authors to solve similar endogeneity problems. In both cases, the authors are clearly taking action to address the endogeneity of income in life-

satisfaction estimations, which is appropriate. However, it is striking and disappointing to see how sensitive the income coefficients are to the method selected. This could support arguments for further discussions and academic work on how best to implement a standard practice for tackling income effects, based on the best available scientific evidence, rather than recalculating them anew in every CIV paper.

Taken together, these results suggest that empirical work seeking externally valid QALY monetary benchmarks should prioritize careful treatment of *losses versus gains*. Whether this pattern reflects only the diminishing marginal utility of income or is further compounded by loss aversion, people disliking income losses more than they value equivalent gains, is not known (Kahneman and Tversky, 1979). In terms of WTP–WTA, the Huang specification estimates a WTP to avoid loss, while the McNamee–Mendolia specification estimates a WTA compensation to bear it. Meta-analyses typically find WTA/WTP ratios in the 1.5–5 range (Hanemann, 1991; Horowitz and McConnell, 2002; O’Brien et al., 1998); our swapped estimates exceed that benchmark by an order of magnitude.

Importantly, both papers adopt entirely conventional income specifications: using a wind-fall dummy directly, or employing a financial shock as an instrument, are both standard strategies for tackling income endogeneity in life-satisfaction equations and both papers cite the same source as justification for their strategy to tackle income (Frijters et al., 2011). The lesson is therefore not that either approach is “wrong,” but that CIV valuations are acutely sensitive to this specification choice.

It is worth noting that other approaches to tackling income endogeneity exist. It is not necessary to estimate the income coefficients in each paper that uses the CIV method. Instead of directly estimating the effect of income on life satisfaction, several studies have relied on previously published income coefficients to bypass some of the known endogeneity issues. How to best do this is not standardized. Published coefficients can be used directly, and regressions constrained to conform to their values. It is also possible to scale income coefficients by the size of the bias reported in published studies that report both uninstrumented

and instrumented results ([Ásgeirsdóttir et al., 2023](#); [Baldursdóttir et al., 2023](#)).

In recent years, the use of wellbeing units (WELLBYs) has also gained traction for the same purpose. Specifically, a WELLBY is defined as “one unit of life satisfaction on a 0–10 scale per person for one year” ([Frijters and Krekel, 2021](#)). It is then assigned a monetary value based on published estimates, thereby avoiding the need to estimate an income effect in each new study directly. This approach has advanced furthest in the UK, where the UK Treasury has published its own valuation of a WELLBY as GBP 13,000 at 2019 prices, with a low-range of GBP 10,000 and a high-range of GBP 16,000 ([HM Treasury, 2021a,b](#)). When compared with our findings, this value aligns most closely with the income coefficients used in [Huang et al. \(2018\)](#). However, there is no consensus in the literature on the optimal approach, and many authors continue to estimate their own income coefficients, as done in both [Huang et al. \(2018\)](#) and [McNamee and Mendolia \(2018\)](#). Similarly, it remains common for researchers to borrow coefficients from established studies rather than deriving them anew (see e.g. [Ásgeirsdóttir et al., 2023](#); [Baldursdóttir et al., 2023](#)).

In the CIV framework, the specification of the income–utility relationship is a critical element, yet it remains a source of considerable modeling heterogeneity across studies. While future research may clarify the best estimation strategy for this component, the literature would already benefit from a degree of methodological standardization to improve comparability of results. By analogy, the cost-effectiveness analysis (CEA) field has formalized such standards: for example, the Second Panel on Cost-Effectiveness in Health and Medicine recommends that all cost-effectiveness studies include a “reference case” analysis using a consistent set of methods to enhance quality and comparability of outcomes ([Sanders et al., 2016](#)). Following this example, could be fruitful, by which researchers report their CIV results under a common reference specification, such as adopting a standard treatment of income, alongside their preferred model, thereby facilitating more straightforward cross-study comparisons without impeding methodological innovation.

Our analyses reveal the primary source of the discrepancy between the results presented

by [Huang et al. \(2018\)](#) and [McNamee and Mendolia \(2018\)](#), namely, the distinct income effects resulting from different methodological approaches. Interestingly, the authors of the two papers appear to have written their manuscripts independently of one another, each expecting to produce novel results and employing methods that can generally be deemed appropriate.

Given the emphasis on novelty within economic publications, one wonders whether both papers would have been produced if one had been published a few years earlier, or whether the authors might then have opted against pursuing something so closely related to an already published paper. Similarly, had one paper been published well before the other, the later authors might have followed the same methods as the earlier paper, and this discrepancy would not have been revealed. This, of course, we will never know.

Econometric rigor has taken center stage in recent years, first through the “credibility revolution,” and more recently through a growing “replication revolution” ([Nosek et al., 2022](#); [Vazire, 2018](#)). Our own detective work underscores why both movements matter: we could replicate each study faithfully, reproducing the original estimates with only minor discrepancies, yet the resulting policy numbers diverged dramatically once the income specification shifted. Crucially, the methods adopted in both papers are well-established ways to address income endogeneity in life-satisfaction equations, which means that “sensible” choices alone are no guarantee of convergent results.

These findings reinforce the value of replication, robustness, and sensitivity checks in their broadest sense, not only rerunning the same code on the same data, but also probing alternative yet defensible specifications. Although we do not explore different institutional contexts or utilize new data sources for the same research question, the importance of such analyses is suggested by our results. Our message might be that while novel research questions are vital, novelty cannot come at the expense of systematic re-examination. Rarely does a single paper deliver a definitive answer; more often, knowledge accumulates point by point, like a pointillist painting, until a coherent picture finally comes into view. Continu-

ous, transparent replications and re-examinations are therefore not a luxury, they are the scaffolding on which reliable economic evidence is built.

References

- Angrist, J. D. and Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30.
- Ásgeirsdóttir, T. L., Harárdóttir, H., and Jónbjarnardóttir, B. (2023). Putting a price on pain: The monetary compensation needed to offset welfare losses due to violence. *Social Science & Medicine*, 336:116268.
- Baldursdóttir, K., McNamee, P., Norton, E. C., and Ásgeirsdóttir, T. L. (2023). Life satisfaction and body mass index: Estimating the monetary value of achieving optimal body weight. *Review of Economics of the Household*, 21(4):1215–1246.
- Brazier, J., Roberts, J., and Deverill, M. (2002). The estimation of a preference-based measure of health from the sf-36. *Journal of Health Economics*, 21(2):271–292.
- Chang, A. C. and Li, P. (2018). Is economics research replicable? sixty published papers from thirteen journals say “often no”. *Critical Finance Review*, 7(2–3):211–255.
- Christensen, G., Miguel, E., and Gneezy, U. (2019). Research transparency in economics. *Journal of Economic Literature*, 57(3):699–747.
- Frijters, P., Johnston, D. W., and Shields, M. A. (2011). Life satisfaction dynamics with quarterly life event data. *The Scandinavian Journal of Economics*, 113(1):190–211.
- Frijters, P. and Krekel, C. (2021). *A Handbook for Wellbeing Policy-Making: History, Theory, Measurement, Implementation, and Examples*. Oxford University Press, Oxford. Accessed January 2025.
- Hanemann, M. W. (1991). Willingness to pay and willingness to accept: How much can they differ? *American Economic Review*, 81(3):635–647. Accessed: 2025-01-11.

- HM Treasury (2021a). Wellbeing discussion paper: Monetisation of life satisfaction effect sizes. Accessed: 2025-01-11.
- HM Treasury (2021b). Wellbeing discussion paper: monetisation of life satisfaction effect sizes: A review of approaches and proposed approach. Retrieved on April 5th 2024.
- Horowitz, J. K. and McConnell, K. E. (2002). A review of wta/wtp studies. *Journal of Environmental Economics and Management*, 44(3):426–447.
- Huang, L., Frijters, P., Dalziel, K., and Clarke, P. (2018). Life satisfaction, qalys, and the monetary value of health. *Social Science & Medicine*, 211:131–136.
- Institute for Replication (2024). I4r: Large-scale replication projects in economics and related fields. <https://i4replication.org>. Accessed 10 May 2025.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, 2(8):e124.
- Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292.
- Lupia, A. and Elman, C. (2014). Openness in political science: Data access and research transparency. *PS: Political Science & Politics*, 47(1):19–42.
- McNamee, P. and Mendolia, S. (2018). Changes in health-related quality of life: A compensating income variation approach. *Applied Economics*, 51(6):639–650.
- Munafò, M. R. et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1:0021.
- Nosek, B. A. et al. (2015). Promoting an open research culture. *Science*, 348(6242):1422–1425.

- Nosek, B. A., Hardwicke, T. E., Moshontz, H., et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73:719–748.
- O’Brien, B. J., Elsworth, B. F., and Calin, A. (1998). Willingness to pay versus willingness to accept in patients with chronic rheumatic disease. *Health Economics*, 7(6):441–451.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.
- Ryen, L. and 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.
- Sanders, G. D., Neumann, P. J., Basu, A., Brock, D. W., Feeny, D., Krahm, M., Kuntz, K. M., Meltzer, D. O., Owens, D. K., Prosser, L. A., Salomon, J. A., Sculpher, M. J., Trikalinos, T. A., Russell, L. B., Siegel, J. E., and Ganiats, T. G. (2016). Recommendations for conduct, methodological practices, and reporting of cost-effectiveness analyses: Second panel on cost-effectiveness in health and medicine. *JAMA*, 316(10):1093–1103.
- Summerfield, M., Garrard, B., Nesa, M., Kamath, R., Macalalad, N., Watson, N., Wilkins, R., and Wooden, M. (2024). Hilda user manual – release 23.
- Tarlov, A. R. (1989). The medical outcomes study. *JAMA*, 262(7):925.
- Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, 13(4):411–417.
- Ware, J. E. and Sherbourne, C. D. (1992). The mos 36-item short-form health survey (sf-36): I. conceptual framework and item selection. *Medical Care*, 30(6):473–483.
- Watson, N. and Wooden, M. (2012). The hilda survey: a case study in the design and development of a successful household panel survey. *Longitudinal and Life Course Studies*.

A Appendix

A.1 Supplementary Figures and Tables

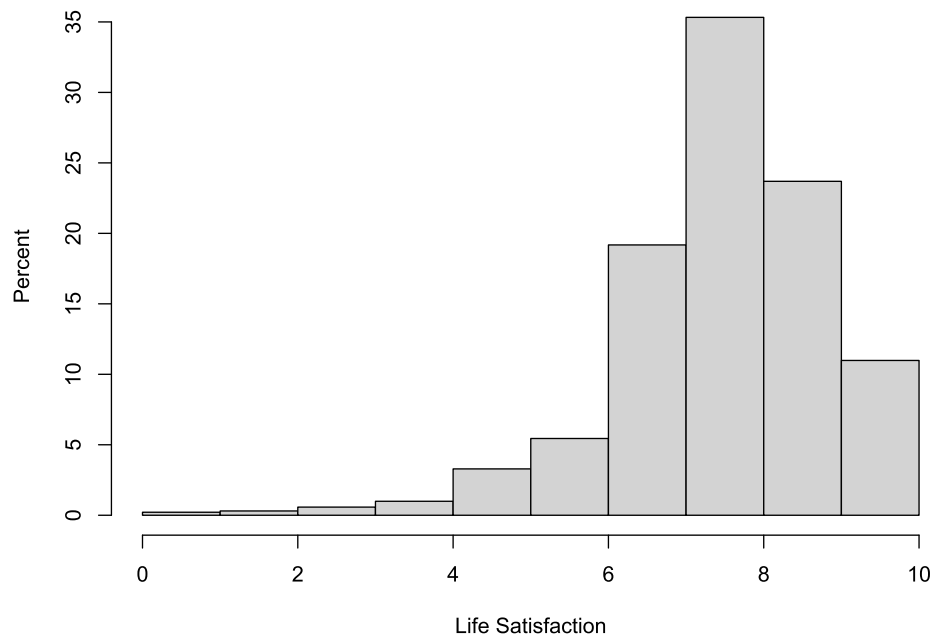


Figure A1: Distribution of life satisfaction

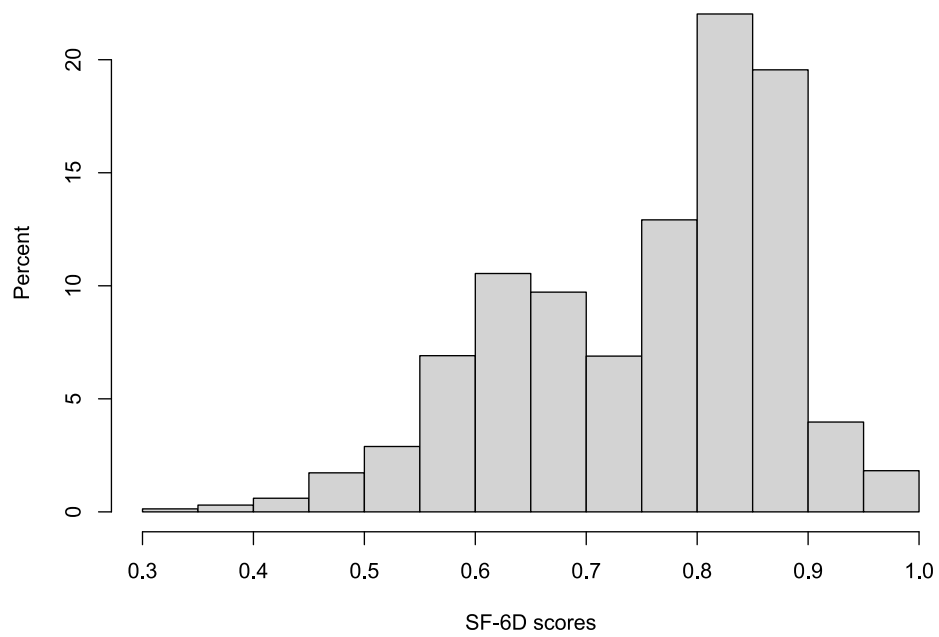


Figure A2: Distribution of SF-6D scores

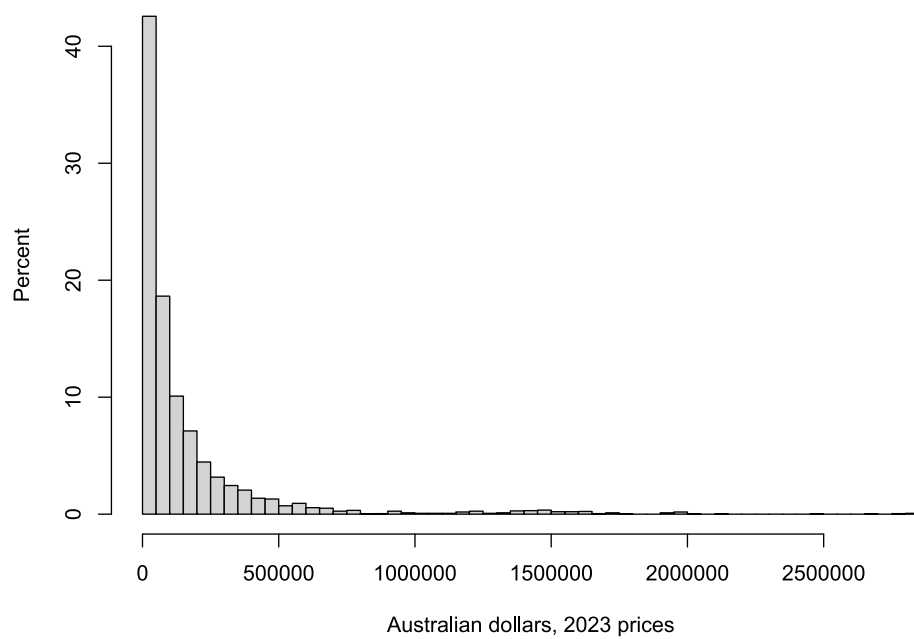


Figure A3: Distribution of gross windfall income

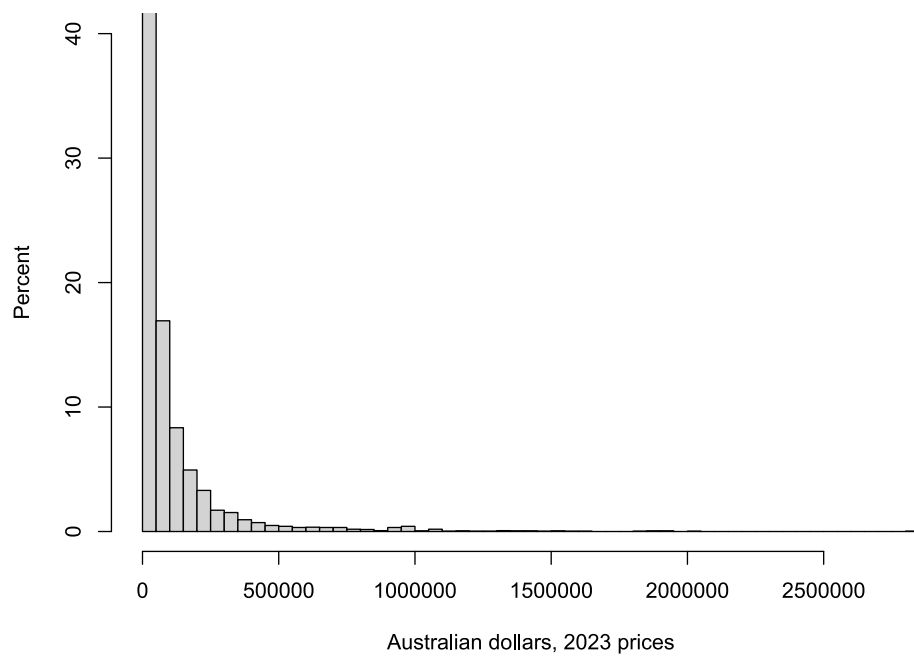


Figure A4: Distribution of equivalized windfall income

Table A1: Replication of Huang et al. (2018) regression coefficients, original waves (2-15) and variables

	Huang et al., no IV (2018) ^a	Huang et al., with IV (2018) ^a	Replication, no IV	Replication with IV
Income in 1000's	0.0006*** (0.0001)	0.080*** (0.010)	0.0005*** (0.00007)	0.064*** (0.007)
Income in 1000's a year ago	0.0004*** (0.0001)	-0.035*** (0.010)	0.0003*** (0.00007)	-0.033*** (0.007)
SF-6D	2.432*** (0.037)	2.258*** (0.134)	2.425*** (0.037)	2.187*** (0.114)
SF-6D a year ago	0.749*** (0.749)	0.778*** (0.136)	0.732*** (0.037)	0.675*** (0.116)
Age	-0.025 (0.015)	0.001 (0.056)	-0.040*** (0.003)	-0.100*** (0.019)
Age squared	0.0004*** (0.000)	0.001*** (0.000)	0.0004*** (0.000)	0.001*** (0.000)
Married/de facto	0.296*** (0.013)	-0.012 (0.066)	0.296*** (0.013)	-0.009 (0.058)
Education	-0.086*** (0.014)	-0.099 (0.054)	-0.090*** (0.015)	-0.100** (0.046)
Leisure capacity	0.042*** (0.012)	0.734*** (0.102)	0.043*** (0.012)	0.719*** (0.091)
Unemployment	-0.170*** (0.019)	-0.225*** (0.068)	-0.161*** (0.019)	-0.143** (0.056)
Number of observations	134,224	106,457	132,425	132,425
R^2			0.053	0.001
Adjusted R^2			-0.125	-0.188

Standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

^aSignificance at 10% level is not reported in Huang et al. (2018) and therefore not reported here.

Table A2: First stage results for IV model replication of Huang et al. (2018) original waves (2-15) and variables

	Income in A\$1000's	Income in A\$1000's, previous year
SF-6D	4.008*** (1.550)	2.409 (1.541)
SF-6D a year ago	3.059** (1.542)	5.459*** (1.534)
Long-term condition	0.774* (0.396)	-0.699* (0.394)
Long-term condition a year ago	-0.663* (1.542)	0.462 (1.534)
Age	2.452*** (0.397)	2.823*** (0.395)
Age squared	-0.014*** (0.001)	-0.018*** (0.001)
Married/de facto	7.162*** (0.539)	4.836*** (0.536)
Education	-0.785 (0.628)	-1.873*** (0.624)
Leisure capacity	-13.945*** (0.505)	-6.420*** (0.503)
Unemployment	0.589 (0.786)	0.686 (0.782)
Financial worsening	-7.046*** (0.783)	-0.060 (0.779)
Financial worsening a year ago	-6.841*** (0.779)	-7.331*** (0.775)
Financial worsening two years ago	-3.898*** (0.718)	-6.818*** (0.714)
Number of observations	132,425	132,425
R^2	0.024	0.018
Adjusted R^2	-0.161	-0.168
F statistic (24, 111387)	112.333***	82.956***

Standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

Year-dummy coefficients are omitted.

Table A3: Replication of monetary values from Huang et al. (2018), original waves (2-15) and variables

	Huang et al., with IV (2018)	Replication with IV
1 QALY 2-year rolling window	42,250	53,392
1 QALY long-run equivalence	67,022	92,395

All values have been converted to 2015 Australian dollars for comparability.

The CIV values for health changes are calculated for the 2-year rolling window, and for the long-run equivalence results, in which concurrent and lagged coefficients are given equal weight.

Table A4: Replication of McNamee and Mendolia (2019) SF-6D changes to specific positions, original waves (1-10) and variables

Changes in SF-6D QALY scores from t_{-1} to t	Model 1 (Mc-Namee and Mendolia, 2019)	Model 2 (Mc-Namee and Mendolia, 2019)	Model 3 (Mc-Namee and Mendolia, 2019)	Model 1 (replication)	Model 2 (replication)	Model 3 (replication)
No change	0.008 (0.009)	0.008 (0.009)	0.006 (0.009)	-0.065*** (0.009)	-0.068*** (0.009)	-0.067*** (0.009)
SF-6D from 0.7 to 0.6	-0.245** (0.025)	-0.246** (0.025)	-0.242** (0.025)	-0.259*** (0.026)	-0.261*** (0.025)	-0.254*** (0.025)
SF-6D from 0.8 to 0.7	-0.120** (0.025)	-0.121** (0.016)	-0.119** (0.016)	-0.146*** (0.017)	-0.149*** (0.017)	-0.146*** (0.017)
SF-6D from 0.9 to 0.8	-0.008 (0.031)	-0.012 (0.012)	-0.011 (0.012)	-0.051*** (0.013)	-0.056*** (0.013)	-0.056*** (0.013)
SF-6D from 1 to 0.9	0.008 (0.031)	0.001 (0.031)	-0.001 (0.032)	-0.033 (0.033)	-0.040 (0.033)	-0.041 (0.033)
SF-6D from > 0.7 to 0.6	-0.244** (0.022)	-0.243** (0.022)	-0.230** (0.022)	-0.273*** (0.023)	-0.272*** (0.022)	-0.257*** (0.022)
SF-6D from > 0.8 to 0.7	-0.112** (0.027)	-0.118** (0.027)	-0.108** (0.026)	-0.140*** (0.028)	-0.143*** (0.027)	-0.135*** (0.027)
SF-6D from > 0.9 to 0.8	0.006 (0.057)	0.003 (0.057)	-0.002 (0.057)	-0.041 (0.059)	-0.042 (0.058)	-0.048 (0.058)
Other neg. changes to < 0.6	-0.627** (0.036)	-0.618** (0.035)	-0.590** (0.035)	-0.578*** (0.038)	-0.576*** (0.038)	-0.549*** (0.037)
Financial improvements	0.121** (0.022)	0.123** (0.022)	0.124** (0.022)	0.135*** (0.022)	0.140*** (0.022)	0.145*** (0.022)
Control for Employment, marital status, and other socio-economic factors	No	Yes	Yes	No	Yes	Yes
Control for personal illness, victim of violence, and other life events	No	No	Yes	No	No	Yes
Observations	83,556	83,530	83,177	80,938	80,938	80,938

Robust standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

Table A5: Replication of McNamee and Mendolia (2019) SF-6D changes of different sizes, original waves (1-10) and variables

Changes in SF-6D QALY scores from t_{-1} to t	Model 1 (Mc-Namee and Mendolia, 2019)	Model 2 (Mc-Namee and Mendolia, 2019)	Model 3 (Mc-Namee and Mendolia, 2019)	Model 1 (replication)	Model 2 (replication)	Model 3 (replication)
No change	-0.030** (0.010)	-0.030** (0.010)	-0.031** (0.010)	-0.030*** (0.010)	-0.033*** (0.010)	-0.035*** (0.010)
Change equal to -0.01	-0.135** (0.030)	-0.135** (0.030)	-0.130** (0.030)	-0.103*** (0.034)	-0.106*** (0.034)	-0.101*** (0.034)
Change equal to -0.02	-0.128** (0.029)	-0.128** (0.028)	-0.130** (0.028)	-0.137*** (0.026)	-0.138*** (0.026)	-0.135*** (0.025)
Change equal to -0.03	-0.148** (0.040)	-0.149** (0.040)	-0.151** (0.040)	-0.169*** (0.042)	-0.168*** (0.042)	-0.164*** (0.042)
Change equal to -0.04	-0.046** (0.012)	-0.047** (0.012)	-0.046** (0.012)	-0.048*** (0.012)	-0.051*** (0.012)	-0.050*** (0.012)
Change equal to -0.05	-0.187** (0.033)	-0.185** (0.033)	-0.186** (0.033)	-0.156*** (0.034)	-0.158*** (0.034)	-0.154*** (0.034)
Change equal to -0.06	-0.151** (0.019)	-0.151** (0.019)	-0.147** (0.019)	-0.144*** (0.019)	-0.146*** (0.019)	-0.144*** (0.019)
Change equal to -0.07	-0.145** (0.035)	-0.143** (0.035)	-0.136** (0.035)	-0.130*** (0.034)	-0.128*** (0.034)	-0.122*** (0.034)
Change equal to -0.08	-0.106** (0.023)	-0.107** (0.023)	-0.102** (0.023)	-0.096*** (0.023)	-0.098*** (0.023)	-0.095*** (0.023)
Change equal to -0.09	-0.202** (0.037)	-0.210** (0.037)	-0.207** (0.037)	-0.172*** (0.042)	-0.179*** (0.041)	-0.175*** (0.041)
Change equal to -0.1	-0.190** (0.030)	-0.198** (0.030)	-0.192** (0.030)	-0.180*** (0.029)	-0.188*** (0.029)	-0.180*** (0.029)
Change < -0.1	-0.244** (0.013)	-0.245** (0.013)	-0.228** (0.013)	-0.244*** (0.013)	-0.245*** (0.013)	-0.230*** (0.013)
Financial improvements	0.123** (0.022)	0.125** (0.022)	0.126** (0.022)	0.137*** (0.022)	0.142*** (0.022)	0.147*** (0.022)
Control for Employment, marital status, and other socio-economic factors	No	Yes	Yes	No	Yes	Yes
Control for personal illness, victim of violence, and other life events	No	No	Yes	No	No	Yes
Observations	83,556	83,530	83,177	80,938	80,938	80,938

Robust standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

Table A6: Replication of monetary values from McNamee and Mendolia (2019) SF-6D changes to specific positions, original waves (1-10) and variables

Changes in SF-6D QALY scores from t_{-1} to t	Model 3 (McNamee and Mendolia, 2019)	Model 3 (replication)
SF-6D from 0.7 to 0.6	208,100	161,911
SF-6D from 0.8 to 0.7	102,449	92,827
SF-6D from 0.9 to 0.8	9,605	35,523
SF-6D from 1 to 0.9	854	26,431
SF-6D from > 0.7 to 0.6	197,428	164,069
SF-6D from > 0.8 to 0.7	92,845	85,989
SF-6D from > 0.9 to 0.8	1,067	30,737
Other neg. changes to < 0.6	506,911	349,614

All amounts are at 2010 price levels for comparability. The valuations use mean windfall income.

Table A7: Replication of monetary values from McNamee and Mendolia (2019) SF-6D changes of different sizes, original waves (1-10) and variables

Changes in SF-6D QALY scores from t_{-1} to t	Model 3 (McNamee and Mendolia, 2019)	Model 3 (replication)
Change equal to -0.01	108,852	63,337
Change equal to -0.02	109,920	84,886
Change equal to -0.03	128,062	103,187
Change equal to -0.04	39,486	31,511
Change equal to -0.05	157,943	96,899
Change equal to -0.06	124,860	90,640
Change equal to -0.07	115,255	76,764
Change equal to -0.08	86,442	59,365
Change equal to -0.09	175,018	109,963
Change equal to -0.1	162,211	113,081
Change < -0.1	193,159	144,345

All amounts are at 2010 price levels for comparability. The valuations use mean windfall income.

Table A8: First stage results for IV model replication of Huang et al. (2018) waves 2-23

	Income in A\$1000's	Income in A\$1000's, previous year
SF-6D	3.668** (1.689)	4.911*** (1.674)
SF-6D a year ago	2.844* (1.685)	5.243*** (1.670)
Long-term condition	-0.014 (0.432)	-0.413 (0.428)
Long-term condition a year ago	-0.958** (0.433)	0.120 (0.429)
Age	3.076*** (0.103)	3.488*** (0.102)
Age squared	-0.019*** (0.001)	-0.022*** (0.001)
Married/de facto	7.517*** (0.545)	5.719*** (0.540)
Education	-2.998*** (0.634)	-3.152*** (0.629)
Leisure capacity	-22.779*** (0.521)	-13.371*** (0.516)
Unemployment	2.002** (0.860)	3.094*** (0.853)
Financial worsening	-8.094*** (0.902)	-1.249 (0.895)
Financial worsening a year ago	-8.685*** (0.893)	-8.256*** (0.885)
Financial worsening two years ago	-4.959*** (0.744)	-8.067*** (0.738)
Number of observations	233,406	233,406
R^2	0.026	0.020
Adjusted R^2	-0.094	-0.101
F statistic (32, 207722)	173.718***	134.219***

Standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

Year-dummy coefficients are omitted.

Table A9: First stage results for IV model replication of McNamee and Mendolia (2019) waves 2-23. SF-6D changes to specific positions.

	Income in A\$1000's	
No change	0.458	(0.317)
SF-6D from 0.7 to 0.6	1.411*	(0.785)
SF-6D from 0.8 to 0.7	0.395	(0.586)
SF-6D from 0.9 to 0.8	0.754	(0.520)
SF-6D from 1 to 0.9	0.939	(1.442)
SF-6D from > 0.7 to 0.6	1.099*	(0.648)
SF-6D from > 0.8 to 0.7	-0.799	(0.905)
SF-6D from > 0.9 to 0.8	0.482	(2.262)
Other neg. changes to < 0.6	-0.291	(0.680)
Financial worsening	-6.631***	(0.973)
Financial worsening, prev. year and earlier	0.070	(1.079)
Unemployed	-8.614***	(0.743)
Out of workforce	-11.723***	(0.688)
Seperated/divorced	-8.579***	(1.601)
Widowed	2.169	(2.515)
Single	-20.985***	(1.030)
House - rent	-16.443***	(0.710)
House - other	-16.422***	(1.301)
Housing tenure - inner city	-4.771***	(1.036)
Housing tenure - outer city	-3.640**	(1.580)
Education - certificate or equivalent	-14.420***	(1.569)
Education - lower than certificate	-20.453***	(1.464)
Have kids aged 0-4, yes or no	-17.219***	(0.543)
Have kids aged 5-9, yes or no	-8.109***	(0.469)
Have kids aged 10-14, yes or no	-9.299***	(0.524)
Life event - Serious personal illness/injury	0.617	(0.546)
Life event - Victim of physical violence	-0.450	(1.090)
Life event - Victim of property crime	-1.821**	(0.724)
Life event - Serious injury/illness of family member	-0.660	(0.437)
Life event - death of a close friend	-0.020	(0.515)
Life event - death of a close relative/family member	5.750***	(0.538)
Life event - death of a spouse or child	6.751***	(2.514)
Number of observations	233,406	
R^2	0.025	
Adjusted R^2	-0.095	
F statistic (32, 207722)	166.978***	

Robust standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

Table A10: First stage results for IV model replication of McNamee and Mendolia (2019) waves 2-23. SF-6D changes of different sizes.

	Income in A\$1000's	
No change	0.771**	(0.383)
Change equal to -0.01	-1.416^*	(0.826)
Change equal to -0.02	1.289	(0.951)
Change equal to -0.03	0.599	(1.135)
Change equal to -0.04	0.343	(0.505)
Change equal to -0.05	0.945	(0.993)
Change equal to -0.06	0.758	(0.633)
Change equal to -0.07	-0.354	(1.023)
Change equal to -0.08	0.331	(0.780)
Change equal to -0.09	-0.460	(0.965)
Change equal to -0.1	0.468	(1.121)
Change < -0.1	0.091	(0.365)
Financial worsening	-6.634^{***}	(0.972)
Financial worsening, prev. year and earlier	0.051	(1.079)
Unemployed	-8.619^{***}	(0.743)
Out of workforce	-11.726^{***}	(0.687)
Seperated/divorced	-8.574^{***}	(1.602)
Widowed	2.170	(2.514)
Single	-20.968^{***}	(1.029)
House - rent	-16.445^{***}	(0.710)
House - other	-16.411^{***}	(1.302)
Housing tenure - inner city	-4.763^{***}	(1.036)
Housing tenure - outer city	-3.626^{**}	(1.580)
Education - certificate or equivalent	-14.421^{***}	(1.569)
Education - lower than certificate	-20.446^{***}	(1.464)
Have kids aged 0-4, yes or no	-17.207^{***}	(0.543)
Have kids aged 5-9, yes or no	-8.106^{***}	(0.469)
Have kids aged 10-14, yes or no	-9.300^{***}	(0.524)
Life event - Serious personal illness/injury	0.654	(0.536)
Life event - Victim of physical violence	-0.442	(1.090)
Life event - Victim of property crime	-1.819^{**}	(0.724)
Life event - Serious injury/illness of family member	-0.647	(0.437)
Life event - death of a close friend	-0.013	(0.515)
Life event - death of a close relative/family member	5.753^{***}	(0.538)
Life event - death of a spouse or child	6.785^{***}	(2.515)
Number of observations	233,406	
R^2	0.025	
Adjusted R^2	-0.095	
F statistic (35, 207719)	152.714***	

Robust standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

Table A11: Replication of Huang et al. (2018) regression coefficients, financial improvement as instrument

	Huang et al., with IV (2018) ^a	Replication with IV
SF-6D	2.258*** (0.134)	2.457*** (0.027)
SF-6D a year ago	0.778*** (0.136)	0.762*** (0.027)
Income in 1000's	0.080*** (0.010)	0.002*** (0.000)
Income in 1000's a year ago	-0.035*** (0.010)	0.001*** (0.000)
Number of observations	106,457	233,406
R^2		0.047
Adjusted R^2		-0.070

Standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

^aSignificance at 10% level is not reported in Huang et al. (2018) and therefore not reported here.

Table A12: Replication of monetary values from Huang et al. (2018), financial improvement as instrument

	Huang et al., with IV (2018)	Replication with IV
1 QALY 2-year rolling window	52,547	998,136
1 QALY long-run equivalence	83,357	1,009,457

All values have been converted to 2023 Australian dollars for comparability.

The CIV value for health changes are calculated for the 2-year rolling window, where changes in the current year weigh twice as much as changes in the year prior and for the long-run equivalence where concurrent and lagged coefficients get equal weight. For sustained changes in the long-term, each year is treated equally.

Table A13: Estimated effects of specific SF-6D QALY scores on life satisfaction, replication of McNamee and Mendolia (2019) with financial worsening instead of windfall income

Changes in SF-6D QALY scores from t_{-1} to t	McNamee and Mendolia (2019)	Replication, financial worsening
No change	0.006 (0.009)	-0.065*** (0.005)
SF-6D from 0.7 to 0.6	-0.242** (0.025)	-0.255*** (0.013)
SF-6D from 0.8 to 0.7	-0.119** (0.016)	-0.125*** (0.009)
SF-6D from 0.9 to 0.8	-0.011 (0.012)	-0.032*** (0.007)
SF-6D from 1 to 0.9	-0.001 (0.032)	0.061*** (0.020)
SF-6D from > 0.7 to 0.6	-0.230** (0.022)	-0.255*** (0.012)
SF-6D from > 0.8 to 0.7	-0.108** (0.026)	-0.119*** (0.015)
SF-6D from > 0.9 to 0.8	-0.002 (0.057)	0.008 (0.034)
Other neg. changes to < 0.6	-0.590** (0.035)	-0.506*** (0.019)
Financial improvements	0.124** (0.022)	-0.525*** (0.023)
Number of observations	83,177	233,406
R^2		0.032
Adjusted R^2		-0.088

Robust standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

Table A14: Estimated effects of different size changes in SF-6D QALY scores on life satisfaction, replication of McNamee and Mendolia (2019) with financial worsening instead of windfall income

Changes in SF-6D QALY scores from t_{-1} to t	McNamee and Mendolia (2019)	Replication, financial worsening
No change	−0.031** (0.010)	−0.016*** (0.006)
Change equal to −0.01	−0.130** (0.030)	−0.157*** (0.019)
Change equal to −0.02	−0.130** (0.028)	−0.088*** (0.014)
Change equal to −0.03	−0.151** (0.040)	−0.148*** (0.023)
Change equal to −0.04	−0.046** (0.012)	−0.025*** (0.007)
Change equal to −0.05	−0.185** (0.033)	−0.139*** (0.019)
Change equal to −0.06	−0.147** (0.019)	−0.131*** (0.011)
Change equal to −0.07	−0.136** (0.035)	−0.139*** (0.019)
Change equal to −0.08	−0.102** (0.023)	−0.106*** (0.013)
Change equal to −0.09	−0.207** (0.037)	−0.189*** (0.022)
Change equal to −0.1	−0.192** (0.030)	−0.139*** (0.016)
Change < −0.1	−0.228** (0.013)	−0.213*** (0.007)
Financial improvements	0.126** (0.022)	−0.538*** (0.023)
Number of observations	83,177	233,406
R^2		0.028
Adjusted R^2		−0.092

Robust standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

Table A15: Monetary estimations for SF-6D changes between different positions, replication of McNamee and Mendolia (2019), financial worsening instead of windfall income

Changes in SF-6D QALY scores from t_{-1} to t	McNamee and Mendolia (2019)	Windfall method with financial worsening
SF-6D from 0.7 to 0.6	290,428	30,645
SF-6D from 0.8 to 0.7	142,979	14,963
SF-6D from 0.9 to 0.8	13,405	3,884
SF-6D from 1 to 0.9	1,192	-7,277
SF-6D from > 0.7 to 0.6	275,534	30,598
SF-6D from > 0.8 to 0.7	129,576	14,249
SF-6D from > 0.9 to 0.8	1,489	-987
Other neg. changes to < 0.6	707,453	60,831
All values have been converted to 2023 Australian dollars for comparability.		

Table A16: Monetary estimations for SF-6D changes of different sizes, replication of McNamee and Mendolia (2019), financial worsening instead of windfall

Changes in SF-6D QALY scores from t_{-1} to t	McNamee and Mendolia (2019)	Windfall method mean
Change equal to -0.01	151,916	18,369
Change equal to -0.02	153,406	10,363
Change equal to -0.03	178,725	17,362
Change equal to -0.04	55,107	2,896
Change equal to -0.05	220,428	16,316
Change equal to -0.06	174,257	15,402
Change equal to -0.07	160,852	16,341
Change equal to -0.08	120,640	12,480
Change equal to -0.09	244,258	22,209
Change equal to -0.1	226,384	16,320
Change < -0.1	269,576	25,033
All values have been converted to 2023 Australian dollars for comparability.		

Table A17: Replication of Huang et al. (2018) regression coefficients, financial improvement as instrument

	Huang et al., with IV (2018) ^a	Replication with financial worsening windfall
SF-6D	2.258*** (0.134)	2.413*** (0.027)
SF-6D a year ago	0.778*** (0.136)	0.740*** (0.026)
Financial worsening	0.080*** (0.010)	−0.474*** (0.014)
Financial worsening a year ago	−0.035*** (0.010)	−0.188*** (0.014)
Number of observations	106,457	233,406
R^2		0.067
Adjusted R^2		−0.049

Standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

^aSignificance at 10% level is not reported in [Huang et al. \(2018\)](#) and therefore not reported here.

Table A18: Replication of monetary values from Huang et al. (2018), financial worsening and no instrument

	Huang et al., with IV (2018)	Replication with financial worsening
1 QALY 2-year rolling window	52,547	309,057
1 QALY long-run equivalence	83,357	300,406

All values have been converted to 2023 Australian dollars for comparability.

The CIV value for health changes are calculated for the 2-year rolling window, where changes in the current year weigh twice as much as changes in the year prior, and for the long-run equivalence where concurrent and lagged coefficients get equal weight.

Table A19: Estimated effects of specific SF-6D QALY scores on life satisfaction, replication of McNamee and Mendolia (2019) with financial worsening instead of windfall income

Changes in SF-6D QALY scores from t_{-1} to t	McNamee and Mendolia (2019)	Replication, financial improvements IV
No change	0.006 (0.009)	-0.066*** (0.005)
SF-6D from 0.7 to 0.6	-0.242** (0.025)	-0.266*** (0.013)
SF-6D from 0.8 to 0.7	-0.119** (0.016)	-0.128*** (0.009)
SF-6D from 0.9 to 0.8	-0.011 (0.012)	-0.033*** (0.007)
SF-6D from 1 to 0.9	-0.001 (0.032)	0.062*** (0.020)
SF-6D from > 0.7 to 0.6	-0.230** (0.022)	-0.264*** (0.012)
SF-6D from > 0.8 to 0.7	-0.108** (0.026)	-0.119*** (0.015)
SF-6D from > 0.9 to 0.8	-0.002 (0.057)	0.006 (0.034)
Other neg. changes to < 0.6	-0.590** (0.035)	-0.522*** (0.019)
Financial improvements	0.124** (0.022)	
Income in 1000's		0.003*** (0.000)
Number of observations	83,177	233,406
R^2		0.016
Adjusted R^2		-0.105

Robust standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

Table A20: Estimated effects of different size changes in SF-6D QALY scores on life satisfaction, replication of McNamee and Mendolia (2019) with financial worsening instead of windfall income

Changes in SF-6D QALY scores from t_{-1} to t	McNamee and Mendolia (2019)	Replication, financial improvements IV
No change	−0.031** (0.010)	−0.017*** (0.006)
Change equal to −0.01	−0.130** (0.030)	−0.153*** (0.019)
Change equal to −0.02	−0.130** (0.028)	−0.095*** (0.014)
Change equal to −0.03	−0.151** (0.040)	−0.154*** (0.023)
Change equal to −0.04	−0.046** (0.012)	−0.025*** (0.007)
Change equal to −0.05	−0.185** (0.033)	−0.143*** (0.019)
Change equal to −0.06	−0.147** (0.019)	−0.136*** (0.011)
Change equal to −0.07	−0.136** (0.035)	−0.141*** (0.019)
Change equal to −0.08	−0.102** (0.023)	−0.110*** (0.013)
Change equal to −0.09	−0.207** (0.037)	−0.197*** (0.022)
Change equal to −0.1	−0.192** (0.030)	−0.143*** (0.017)
Change < −0.1	−0.228** (0.013)	−0.219*** (0.007)
Financial improvements	0.126** (0.022)	
Income in 1000's		0.003*** (0.000)
Number of observations	83,177	233,406
R^2		0.013
Adjusted R^2		−0.110

Robust standard errors in brackets. Significance levels: * is significant at 10% level, ** at 5% level and *** at 1%.

Table A21: Monetary estimattions for SF-6D changes between different positions, replication of McNamee and Mendolia (2019), financial improvements as IV

Changes in SF-6D QALY scores from t_{-1} to t	McNamee and Mendolia (2019)	financial improvements as IV
SF-6D from 0.7 to 0.6	290,428	95,273
SF-6D from 0.8 to 0.7	142,979	45,873
SF-6D from 0.9 to 0.8	13,405	11,909
SF-6D from 1 to 0.9	1,192	-22,108
SF-6D from > 0.7 to 0.6	275,534	94,572
SF-6D from > 0.8 to 0.7	129,576	42,556
SF-6D from > 0.9 to 0.8	1,489	-2,171
Other neg. changes to < 0.6	707,453	187,239
All values have been converted to 2023 Australian dollars for comparability.		

Table A22: Monetary estimattions for SF-6D changes of different sizes, replication of McNamee and Mendolia (2019), financial improvements as IV

Changes in SF-6D QALY scores from t_{-1} to t	McNamee and Mendolia (2019)	financial improvements as IV
Change equal to -0.01	151,916	54,303
Change equal to -0.02	153,406	33,628
Change equal to -0.03	178,725	54,417
Change equal to -0.04	55,107	8,795
Change equal to -0.05	220,428	50,836
Change equal to -0.06	174,257	48,250
Change equal to -0.07	160,852	50,119
Change equal to -0.08	120,640	38,998
Change equal to -0.09	244,258	69,756
Change equal to -0.1	226,384	50,634
Change < -0.1	269,576	77,585
All values have been converted to 2023 Australian dollars for comparability.		

A.2 Deriving CIV specifications of different forms

The single-period benchmark. Start with a standard life-satisfaction regression in which health (H), income (Y) and other controls (X) enter linearly

$$\text{LS} = \alpha + \beta_1 H + \beta_2 Y + \beta_3 X + \varepsilon, \quad (\text{A1})$$

and equate the utility of a *better* health state H_0 with the utility of a *worse* health state H_1 that is offset by additional income CIV:

$$\alpha + \beta_1 H_0 + \beta_2 Y + \beta_3 X = \alpha + \beta_1 H_1 + \beta_2 (Y + \text{CIV}) + \beta_3 X. \quad (\text{A2})$$

Because everything else cancels, we are left with $0 = \beta_1 (H_0 - H_1) + \beta_2 \text{CIV}$. For a one-unit change in health ($H_0 - H_1 = 1$). In this case solving for CIV gives

$$\text{CIV} = -\frac{\beta_1}{\beta_2}. \quad (\text{A3})$$

Windfall-dummy income. If income is measured as a dummy for a financial windfall, as in [McNamee and Mendolia \(2018\)](#), the trade-off is expressed in *units* of that dummy. Multiplying by the mean size of the windfall (\bar{Y}_{wind}) converts the ratio to dollars and the CIV formulation becomes

$$\text{CIV} = -\frac{\beta_1}{\beta_2} \bar{Y}_{\text{wind}}. \quad (\text{A4})$$

Log-income variants. Many papers assume diminishing marginal utility and estimate $\ln Y$. In that case the compensating variation becomes

$$\text{CIV} = \bar{Y} \left[\exp(-\beta_1/\beta_2) - 1 \right]. \quad (\text{A5})$$

Huang et al. (2018) formulations. Below we unpack the Huang et al. (2018) formulas, which some find puzzling because they average life satisfaction over *two* adjacent years (“2-year rolling window”) and, separately, over a steady-state (“long-run equivalence”) horizon.

Estimation equation. Simplifying to the health and income terms that matter for CIV, their fixed-effects regression is

$$\text{LS}_{it} = \alpha + b_0 \text{SF}_{it} + b_1 \text{SF}_{i,t-1} + d_0 Y_{it} + d_1 Y_{i,t-1} + \dots \quad (\text{A6})$$

The single lag ($t - 1$) can be seen as capturing one year of *adaptation* (past health) and one year of *habit* (past income).

What is being priced by Huang et al. (2018)?

- **Two-year window:** a health shock that *started last year and persists this year*. Hence current health hurts life satisfaction in *both* years (b_0), while the lagged effect (b_1) appears only in the second year.
- **Compensation path:** Huang assume the same annual top-up CIV is paid in Year 1 *and* Year 2. Therefore the current-income coefficient d_0 enters twice (one for each year) and the lagged coefficient d_1 enters once (felt only in Year 2).

Deriving the 2-year rolling-window CIV. Stack the two annual LS equations (details in Appendix text) and impose $\text{LS}_{t-1,t}^{\text{good}} = \text{LS}_{t-1,t}^{\text{bad+comp}}$. Collecting terms gives

$$2b_0 + b_1 = (2d_0 + d_1) \text{CIV},$$

so that

$$\text{CIV}_{2\text{-yr}} = \frac{2b_0 + b_1}{2d_0 + d_1}. \quad (\text{A7})$$

Interpretation: pay the affected individual $CIV_{2\text{-yr}}$ in *each* of the next two years and her summed life satisfaction over those two years is the same as if she had remained healthy.

Deriving the long-run equivalent CIV. In steady state the health loss and the income supplement have been in place for more than one year, so current and lagged values coincide. Setting per-period utilities equal yields

$$CIV_{\text{long-run}} = \frac{b_0 + b_1}{d_0 + d_1}. \quad (\text{A8})$$

Interpretation: $CIV_{\text{long-run}}$ is the annual stipend that must be paid *forever* to offset a permanent one-unit drop in SF-6D.

Explaining the weights. The apparently “2–1” weighting in (A7) is mechanical: the current-year coefficient (b_0) affects both Year 1 and Year 2, the lagged coefficient (b_1) only Year 2. The same counting logic applies to the income coefficients. If Huang had included two lags, the numerators and denominators would have three terms; if they had assumed a *one-off* payment in Year 1 only, the denominator would be $d_0 + d_1$ instead of $2d_0 + d_1$.

Relation to the broader literature. McNamee & Mendolia use the more commonly used Equation (A3) (or (A4) depending on how you look at it) because their specification has no lags, implicitly valuing a *concurrent* health shock with a matching one-off income change. Although unusual, Huang’s two formulas, therefore, do not contradict the “textbook” CIV; they value *different timelines*. In short, Huang’s $(2b_0 + b_1)$ and $(b_0 + b_1)$ numerators, and their matching income denominators, are nothing more than transparent bookkeeping of how many times each coefficient influences the summed (or steady-state) utility being equated. Understanding that timeline reveals that the two studies are *not* directly comparable until the income treatment and the implicit time horizon are made consistent. Although this is not what appears to drive the large differences in this study, a consistent theoretical foundation

for this literature could be helpful.