The European Journal of Health Economics
https://doi.org/10.1007/s10198-026-01928-1

ORIGINAL PAPER

Do we agree on health state outcomes? A review of measurement agreement between time-trade off and other utility measures

Péter György Balázs[1] · **Valentin Brodszky**[1]

Received: 18 September 2025 / Accepted: 2 April 2026
© The Author(s) 2026

Abstract

Aim The aim of this review is to investigate the measurement agreement of time trade-off (TTO) and direct/indirect health utility measurements methods. Discrepancies have been reported between utility elicitation methods, thus the study objective was to collect all empirical studies that investigated measurement by Bland–Altman analysis (BA) and estimate overall means differences.

Methods A systematic literature review was performed in 2025 April, on three online databases (PubMed, Web of Science, Cochrane) following PRISMA guideline to synthetize (1) original, (2) English language studies, (3) investigating measurement agreement between TTO and other direct and indirect utility measures (4) by BA. Bayesian meta-analysis was performed to estimate overall mean difference and heterogeneity between measures.

Results Overall, $n=402$ records were found, $n=41$ assessed in full text and finally $n=12$ studies were included into the synthesis. The studies covered nine different diseases, the mean TTO utility scores ranged between 0.96 (patient experienced myopia) and 0.42 (patient experienced colorectal cancer). The pooled means differences between the TTO and direct/indirect measures was small (-0.01 and 0.01), however the 95% lower–upper confidence intervals warns that mean estimates can deviate by 0.1 to 0.2. Moderate study heterogeneity ($\tau=0.04$ and $\tau=0.13$) also points on considerably varying utility results study-to-study.

Conclusion Between TTO and other direct/indirect utility measures our review found small mean differences, however significant between-study heterogeneity is indicating inconsistent measurement agreement. Currently, whether discrepancies arise from valuation technique, instrument properties, or study context remained undiscovered.

Keywords Health-related quality of life · Health state utilities · Direct and indirect utility measurement · Bland-Altman measurement agreement · Systematic literature review · Bayesian meta-analysis

Introduction

Health policy interventions outcomes are often assessed by health economic evaluations to support decision making and help better allocation of resources. Evaluations like cost-utility or cost-effectiveness analysis apply quality adjusted life-years index (QALY) to measure intervention effect (health gain/benefit) [1]. The QALY has two components, (1) quality of life multiplied by (2) quantity of life, where one year in full health equals one QALY.

The quantity of life is given by the life expectancy, while quality of life is quantified through health state utility [2]. The utility value may refer to better-than-dead (BTD) health states, anchored between '1' expressing full health and '0' equal to death. Negative utility values represent worse-than-dead (WTD) health states [3].

Health utilities are core inputs for cost-effectiveness evaluations as directly determine QALY estimates [4]. Differences in utility outcomes not only limit comparability, but even small differences in utility outcomes can influence financing decisions, particularly when cost/QALY results are close to the threshold limit (maximizing the amount that a country is able/willing to spent for health gain) [5].

Health utility can be calculated using direct or indirect methods. *Direct utility measurements* like time trade-off (TTO) and standard gamble (SG) are designed to elicit

✉ Péter György Balázs
peter.balazs@uni-corvinus.hu

[1] Department of Health Policy, Corvinus University of Budapest, Fővám Tér 08, 1093 Budapest, Hungary

Published online: 19 June 2026

preferences, through a choice-based task, where alternatives are offered for the respondents. For example, in a TTO task the respondent has to choose between living 10 years in a described imperfect health state or living a shorter period 10-x years in perfect health. The utility of the given health state is elicited from the point of indifference (U=(10-x)/10), where the two alternatives represent the same value for the respondent [6]. TTO is more often used than other direct utility measures and has an often-mentioned benefit of being able to value BTD and WTD health states better [7]. TTO has diverse methodological variation, as the offered timeframe, description of health state, anchor of perfect health, and iteration process can be customized [8]. The SG task works similarly to TTO, but with risk of death [9]. The visual analogue scale (VAS) is debated as a direct utility elicitation, as asking the respondent to rate the given health state on a 0 (worst) to 100 (best) scale incorporates no choice between alternatives [10].

Indirect utility is measured by multi-attribute utility questionnaires (often referred as preference-accompanied or patient reported outcome measures), where respondents assess physiological/psychological/social health dimensions on the designed rating scale [11]. The instrument score is transformed into utility using societal preference weights (also called value sets or tariffs), based on the results of direct measurement method [12]. While indirect utility measures are less time consuming and cognitively less demanding, limit the evaluation by the covered health-related quality of life (HRQoL) domains of generic (e.g. EQ-5D-5L) or disease specific instruments (such as European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire Core 30) [13, 14].

Discrepancies within elicitation methods may stem from multiple reasons, such as the difference in the task attributes [8, 15], or alterations in the psychometric properties of the HRQoL instrument [16, 17], disease/population/demography specific attributes (like, younger women nurturing children, older cancer patients facing decreased lifespan), but scatter in utility outcomes might be further amplified by the widespread use of national value sets with various valuation protocols leading to non-interchangeable utility estimates for identical health states [18]. Bland–Altman (BA) analysis was designed to evaluate agreement between two measurement approaches by quantifying systematic bias (e.g. difference in mean values) and limits of agreement, and over the past four decades it has been widely adopted in health research and outcome measurement studies [19, 20]. The aim of this review is to investigate the measurement agreement of TTO with direct and indirect utility elicitation methods among empirical studies that conducted BA analysis. Secondary objective was to estimate the magnitude of means differences with meta-analysis and compare agreement parameters.

Methods

Search strategy

A systematic literature search was performed on 15th of April 2025 in three online databases: (1) PubMed, (2) Web of Science, (3) Cochrane library. A keyword-based search strategy was developed, the search followed PRISMA guideline [21]. No language and publication date or type filters were applied. (See the detailed search strategy in the Supplementary Material).

Study selection

Articles were screened based on title and abstract according to pre-defined criteria: (1) English language, (2) original studies, (3) reporting TTO measured utilities along (4) Bland–Altman agreement parameters to pool studies investigating measurement agreement between the TTO and any other direct and/or indirect health utility measurement methods. The scope of the review was limited on studies which conducted BA analysis estimating agreement bias and limit of agreement instead of association-based measures or comparison of instrument properties, thus studies relying on concordance, correlation, mapping approaches were excluded [22]. During the full-text assessment, a manual search was conducted in the reference lists, to find further potentially relevant articles. The screening and eligibility assessment were carried out by PB & VB independently, disagreements were resolved through consensus by authors. The references were downloaded and managed using Endnote X8 than exported into csv file for screening.

Data extraction

Study characteristics related data (publication year, country, study design, sample population, sample size, proportion of female, investigated health state/diagnosis, data collection mode and method), TTO method related attributes (type, timeframe, evaluated health state, utility anchor at state '1', the iteration process, number of TTO responses) and the agreement parameters between utility measurements (means and standard deviation, correlations, Bland Altman mean-differences and limits of agreement) were extracted from each publication [23]. For studies with multiple time-point utility estimates, only the baseline comparison was considered.

Utility measurement methods and instruments in review

The following section introduces briefly the five generic and three disease specific indirect utility measurement instruments covered in this review. The *EQ-5D* stands the most widely and most often used generic HRQoL measurement instrument, consisting of a five-dimensional (mobility, self-care, usual activities, pain/discomfort, anxiety/depression) health state description rated on three (3L) or five (5L) levels. There are more than 30 national tariffs, to elicit indirect utility scores according to the preferences of the adequate population [24]. *The Short-Form 6-Dimensions* (SF-6D) was designed to be a simplified generic health descriptive system, covering physical functioning, role limitation, social functioning, pain, mental health and vitality domains, rated on 4 or 6 levels depending on the dimension [25]. The *15D* questionnaire is also a widely adopted generic HRQoL measurement instrument, covering physical, psychological and social dimensions of health with 15 questions rated on 1–5 response levels [26]. The *Patient-Reported Outcomes Measurement Information System* (PROMIS) Preference score (PROPr) is a new, generic, preference-based HRQoL measure, mostly popular in the country of origin (US) [27]. *Assessment of Quality-of-Life* (AQoL-7D) is a 26-item health descriptive system, covering seven separate domains (independent living, relationships, mental health, coping, pain, senses, vision), rated on a 1–5 scale [28]. The last domain of the AQoL-7D instrument is a bolt-on of a separate instrument, called *Vision-related Quality of Life Index* (VisQol), designed for assessing outcomes of healthcare interventions among visually impaired populations. VisQol consists of six items covering six health domains, rated on five response levels [29]. Both questionnaires were developed in Australia, by the same research team providing the only available TTO-based value set for utility preference weighting [28, 30]. The six-item *Visual Function Questionnaire–Utility Index* was also developed for measuring vision-related functioning and well-being of patients to be used in policy decision-supporting health evaluations [31]. The *Dermatology Life Quality Index* (DLQI) is the most frequently used skin-disease specific HRQoL questionnaire. The DLQI covers six dimensions of health (symptoms and feelings, daily activities, leisure, work and school, personal relationships and treatment), with 10 items, each item is rated at 0–3 severity levels [32].

Quality assessment of the included studies

The quality assessment of the included studies was performed according to the Joanna Briggs Institute's (JBI) Critical Appraisal Checklist for Analytical Cross Sectional Studies tool [33]. Five of the eight instruments having relevance for the included studies were selected for the rating: 1) clearly defined inclusion criteria for the sample population, 2) detailed description of the study subjects and settings, 3) measuring the assessed condition using objective criteria, 4) measuring the outcomes in a valid and reliable way, 5) appropriateness of the applied statistical methods. The five items were rated as yes (1), unclear/not reported (0), no (−1).

Meta-analysis method

Two meta-analysis and regression model were designed to evaluate the agreement between 1) TTO and direct measures and 2) between TTO and indirect measurement instruments. Bayesian-meta-analysis with a random-effect model was used to estimate the overall mean difference and between-study heterogeneity (the standard deviation of mean differences) between TTO and direct/indirect utility measurement methods. The priors (of mean difference) were set conservatively to be centered at 0.2 with a standard deviation (SD) of 0.1 assuming small mean differences and heterogeneity, according to previous evidence on discrepancies between direct and indirect health utility estimates [34]. The effect of five binary coded predictors (*data collection*: 1= interview/0= self-administrated; *TTO type*: 1= conventional/0= composite; *population:* 1= patient/0= general; *anchor state*: 1= full health/0= else; *disease*: 1= cancer/0= else) on utility mean differences were estimated using Bayesian meta-regression. Model diagnostics included convergence and fit metrics along with effective sample size check to be able to report reliable posterior estimates. Along with the overall mean differences and corresponding SD, the 95% credible intervals (CI) were reported to display the magnitude and lower–upper bound of the posterior (true) estimates. Sensitivity analysis was performed by varying priors (by ± 0.05) for both the mean difference and between-study heterogeneity parameters to assess robustness of the meta-analysis results. The meta-analysis was conducted using RStudio (v4.3.3.) dplyr, ggplot2 and brms packages [35].

Results

Study selection

Overall $n = 402$ records were identified on the three databases, after removing duplicates $n = 284$ were screened according to title/abstract, $n = 41$ articles were assessed full-text, where $n = 10$ studies met the inclusion criteria, and two further studies were found to be eligible among the reference list of assessed articles. During full-text assessment,

most studies ($n=16$) were excluded for i) not measuring agreement between TTO and other method [36–51], ii) $n=14$ did not apply BA analysis for measurement agreement [52–65] and one paper was not available online [66]. All $n=12$ studies were included into the meta-analysis measurement agreement among TTO vs other utility estimates [67–78]. (Fig. 1 – PRISMA flowchart of the search process).

Study characteristics

Included studies were published between 1998 and 2025, conducted in seven different countries (China, Hungary, India, Iran, Norway, South Korea, Thailand), with sample sizes ranging between 57 to 765. Ten studies derived patients' utilities, two examined general population reported utility values. The research covered nine different diseases of patients (breast cancer, colorectal cancer, chronic obstructive pulmonary disease, dermatological diseases, epilepsy, locally advanced cervical cancer, myopia, pemphigus, visual impairment), and various health states among the general population. All studies were observational, eleven followed cross-sectional design, one was longitudinal. Majority of the studies ($N=8$) used face-to-face interviews for data collection, three followed (online or paper based) self-completed data collection. The longitudinal study combined face-to-face interview for direct utility assessment tasks with paper based self-completed questionnaires for indirect utility measurement. Sample population mean ages ranged widely (25.2–57.0); the proportion of female varied between 23.5–100% (Table 1).

Time trade-off method attributes

Ten studies used conventional TTO type, two applied cTTO task. The 10-year timeframe was set more frequently ($N=6$), than subjective life expectancy ($N=4$) in conventional TTO tasks. One of the cTTO tasks applied a 10-years in BTD and 10+10 years in WTD scenario, the other study had 4-year BTD and 4+4-years WTD timeframe. All but one studies evaluated self-experienced health states, described as respondents own-current condition. One study investigated 46 different hypothetical health states, described by

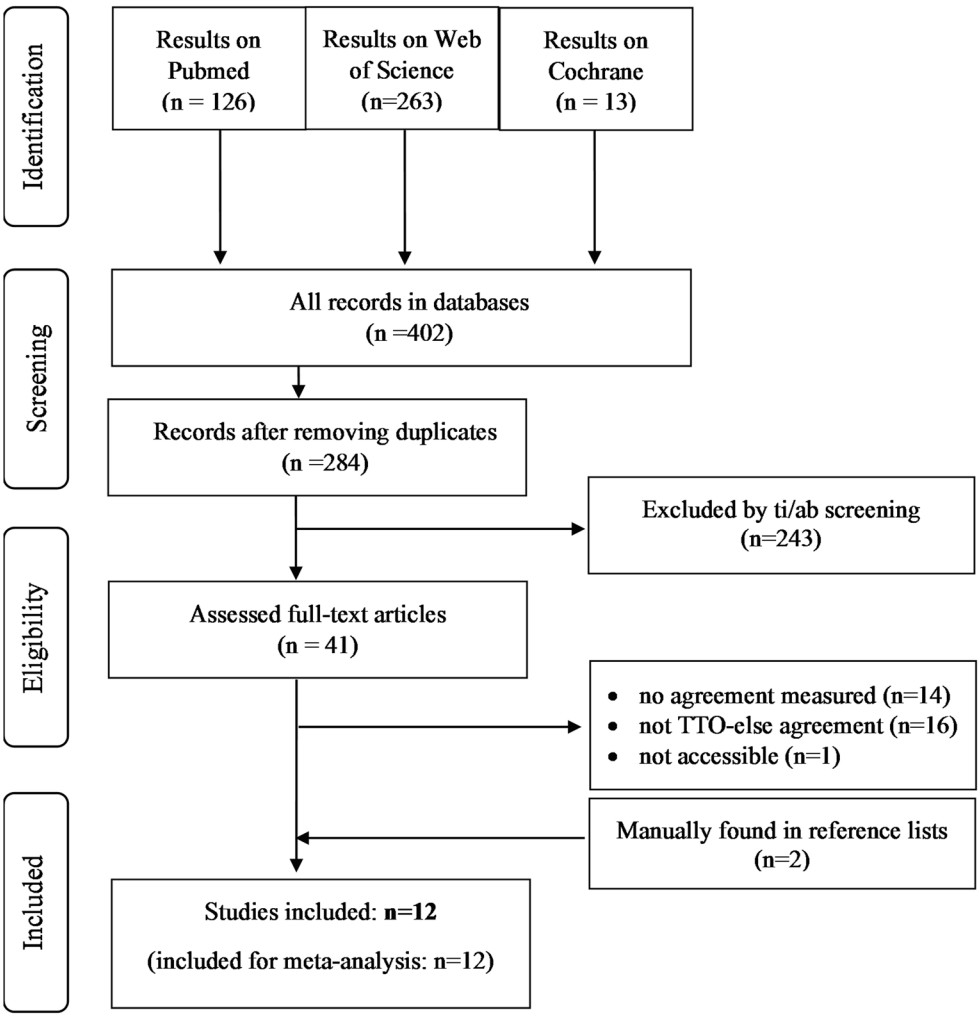

Fig. 1 PRISMA flowchart of the literature search process

Table 1 Study characteristics

Study	Publ. year	Country	Study type (design)	Sample population	Investigated health state*	Sample size (n)	Mean age (SD)	Proportion of female n (%)	Data collection method (mode)
Balázs et al. 2025	2025	Hungary	cross-sectional (observational)	patients	chronic dermatological disease (ICD-10 diagnosed)	765	41.5 (16.2)	362 (47.3%)	self-completed (paper based)
Dou et al. 2024	2024	China	cross-sectional (observational)	patients	myopia (LASIG surgery patients)	477	25.2 (6.0)	237 (49.7%)	interview (face-to-face)
Szabó et al. 2024	2024	Hungary	cross-sectional (observational)	general population	chronic skin condition (self-admitted)	120	median: 51	73 (60.8%)	self-completed (online survey)
Katanyoo et al. 2021	2021	Thailand	cross-sectional (observational)	patients	locally advanced cervical cancer (physician verified by FIGO)	194	53.4 (11.4)	194 (100%)	interview (face-to-face)
Yousefi et al. 2019	2019	Iran	cross-sectional (observational)	patients	colorectal cancer (in/outpatient care visitors)	223	55.2 (13.0)	71 (31.8%)	interview (face-to-face)
Li et al. 2018	2018	China	cross-sectional (observational)	patients	breast cancer (treated inpatients)	608	48.0 (9.6)	608 (100%)	interview (face-to-face)
Liu et al. 2018	2018	China	cross-sectional (observational)	patients	psoriasis vulgaris (outpatient care visitors)	343	39.4 (12.9)	105 (30.6%)	interview (face-to-face)
Kim et al. 2017	2017	South Korea	cross-sectional (observational)	general population	various health states (described with EQ-5D)	500	43.5 (14.5)	251 (50.2%)	interview (face-to-face)
Li et al. 2014	2014	China	cross-sectional (observational)	patients	myopia (scheduled refractive surgery participants)	442	23.7 (5.3)	185 (41.9%)	interview (face-to-face)
Gothwal et al. 2013	2013	India	cross-sectional (observational)	patients	visual impairment (physician verified)	349	43.4 (17.8)	82 (23.5%)	interview (face-to-face)
Stavem et al. 1998	1998	Norway	cross-sectional (observational)	patients	epilepsy (outpatients)	57	43.7 (11.7)	33 (57.9%)	self-completed (paper based)
Stavem et al. 1999	1999	Norway	longitudinal (observational)	patients	COPD (outpatients)	59	57.0 (9.1)	25 (42.4%)	interview (face-to-face) and self-completed (paper based)

* COPD = Chronic obstructive pulmonary disease, ICD-10 = International Classification of Diseases (version 10), FIGO = International Federation of Gynecology and Obstetrics

EQ-5D-5L vignettes. The anchor state for 'utility = 1' varied as follows: full health (N = 7); perfect health (N = 2); perfect vision (N = 1); full visual function (N = 1); restored vision (N = 1). The iteration processes followed most often indifference in one answer (N = 5) or top-down (N = 2) titration. Bisectional, two step and ping-pong iteration was applied by one study each, while two studies did not report the mode of iteration process (Table 2).

Measurement agreement of TTO with health utility measures

The mean TTO utility ranged between 0.96 (patient's experienced myopia) to 0.42 (patient's health in colorectal cancer) among the investigated health states. Overall, 28 comparisons were made between TTO and direct/indirect utility measurements: SG (n = 6), VAS (n = 2) and EQ-5D-5L/3L (n = 6 and 2), 15D (n = 2), AQoL-7D (n = 1), SF-6D (n = 1), VFQ-UI (n = 1), VisQoL (n = 1). Among direct utility measure comparisons, the TTO mean was higher only in three out of eight cases (38%), while compared to indirect utility measurements in thirteen out of twenty cases (65%) the mean TTO score was higher Figs. 2 and 3.

The internal consistency of TTO with direct utility measures (ICC = 0.11–0.87) was weak to strong and weak to moderate with indirect measures (ICC = 0.12–0.50). Nine out of twelve studies investigating measurement agreement has stated no or poor agreement between TTO and direct/indirect measures, half undeclared advice on measurement instrument use (Table 3).

Study quality assessment results

The twelve included studies were homogeneously representing excellent quality. Based on the five JBI rating items of (1) inclusion criteria description, (2) study setting details, (3) measurement appropriateness of the utility elicitation, (4) outcomes validity/reliability properties and (5) correct statistical analysis, ten studies had max rating score. Two studies missed to report the applied iteration process during the TTO task, which might limit reproduction of results as the titration mode has reported effects on utility outcomes (Table 4).

Table 2 Time trade of (TTO) task methodological attributes

Study	TTO method	Timeframe	Evaluated health state (disease)	Health state description	Description of '1' utility state	Iteration process**	N of TTO responses
Balázs et al. 2025	conventional	10-year	self-experienced (atopic dermatitis, hidradenitis s., pemphigus, psoriasis)	own-current	full health	top-down	730
Dou et al. 2024	conventional	subjective life expectancy	self-experienced (myopia)	own-current	restored vision	nr	477
Szabó et al. 2024	conventional	10-year	self-experienced (any skin condition)	own-current	full health	top-down	120
Katanyoo et al. 2021	conventional	10-year	self-experienced (cervical cancer)	own-current	perfect health	indifference in one answer	194
Yousefi et al. 2019	composite	4-years in BTD and 4 + 4-years in WTD	self-experienced (colorectal cancer)	own-current	full health	ping-pong with randomized starting	223
Li et al. 2018	conventional	10-year	self-experienced (breast cancer)	own-current	full health	indifference in one answer	608
Liu et al. 2018	conventional	subjective life expectancy	self-experienced (psoriasis vulgaris)	own-current	perfect health	two step iteration	343
Kim et al. 2017	composite	10-years in BTD and 10 + 10 years in WTD	hypothetical (46 health state combinations)	vignettes: based on EQ-5D-5L states	full health	bisectional, starting with midpoint	500
Li et al. 2014	conventional	subjective life expectancy	self-experienced (myopia)	own-current	perfect vision	nr	442
Gothwal et al. 2013	conventional	subjective life expectancy	self-experienced (visual impairment)	own-current	full visual function	indifference in one answer	251
Stavem et al. 1998	conventional	10-year	self-experienced (epilepsy)	own-current	full health	indifference in one answer	57
Stavem et al. 1999	conventional	10-year	self-experienced (COPD)	own-current	full health	indifference in one answer	59

*BTD = Better than dead (health state), WTD = worse than dead (health state)

**nr = not reported

Meta-analysis of utility measurements agreement

Across 8 comparisons of TTO and direct utility measures the pooled mean difference was −0.01 with a 95% CI of −0.04–0.02 indicating the absence of systemic bias between measurement methods. The between-study heterogeneity ($\tau = 0.04$) indicates slightly varying mean differences among the eight studies, the true mean difference was < 0.1. The good model convergence and fit (Rhat = 1.0, good Bulk and Tail values) also supports that the true (posterior) overall mean difference between TTO vs SG and VAS is small. The Bayesian meta-regression results on factors that might impact TTO-direct utility mean difference showed, that none of the evaluated study-level covariates (data collection mode, TTO type, population, anchor state, disease) had meaningful associations with mean differences.

The pooled mean difference of 20 TTO-indirect utility measurement pairs was close to zero, but the range of true mean difference is rather large (95%CI: −0.08–0.09), with moderately robust between-study variability ($\tau = 0.13$)), that express bias in agreement across studies, since utility mean scores can differ by ± 0.1–0.2 depending on the measurement method. The meta-regression results showed no effect of mean differences among TTO and indirect measures, considering the included factors. Sensitivity analyses by smaller/larger priors confirmed the meta-analysis outcomes, the pooled mean differences and between-study heterogeneity of TTO-direct and TTO-indirect comparisons were robust to prior specification (Table 5 and Fig. 4).

Discussion

Disagreement between direct and indirect utility measurement methods introduces uncertainty to economic evaluations by limited comparability of quality of life across health states and thus can alter policy decisions [4, 79]. Our review aimed to synthetize empirical evidence on the extent of measurement agreement between TTO and direct/indirect utility methods essential for supporting resource allocation decisions through more sensitive health economic evaluations. According to the pooled results of

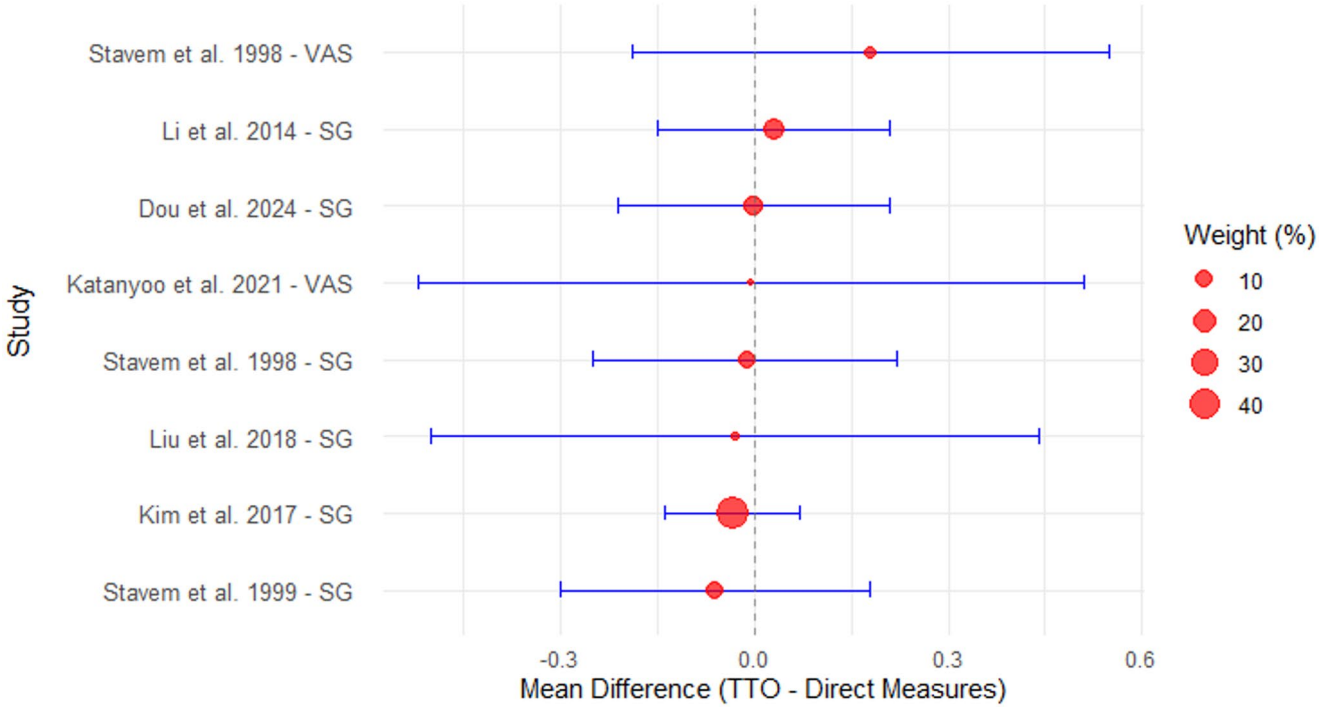

Fig. 2 Forest plot of overall mean difference between TTO and direct utility measurements. * The meta-analysis pooling means differences was close to zero with a 95%CI of −0.04 to 0.03 indicating no systematic bias between TTO and direct utility measures. TTO tends to result in slightly lower utility scores

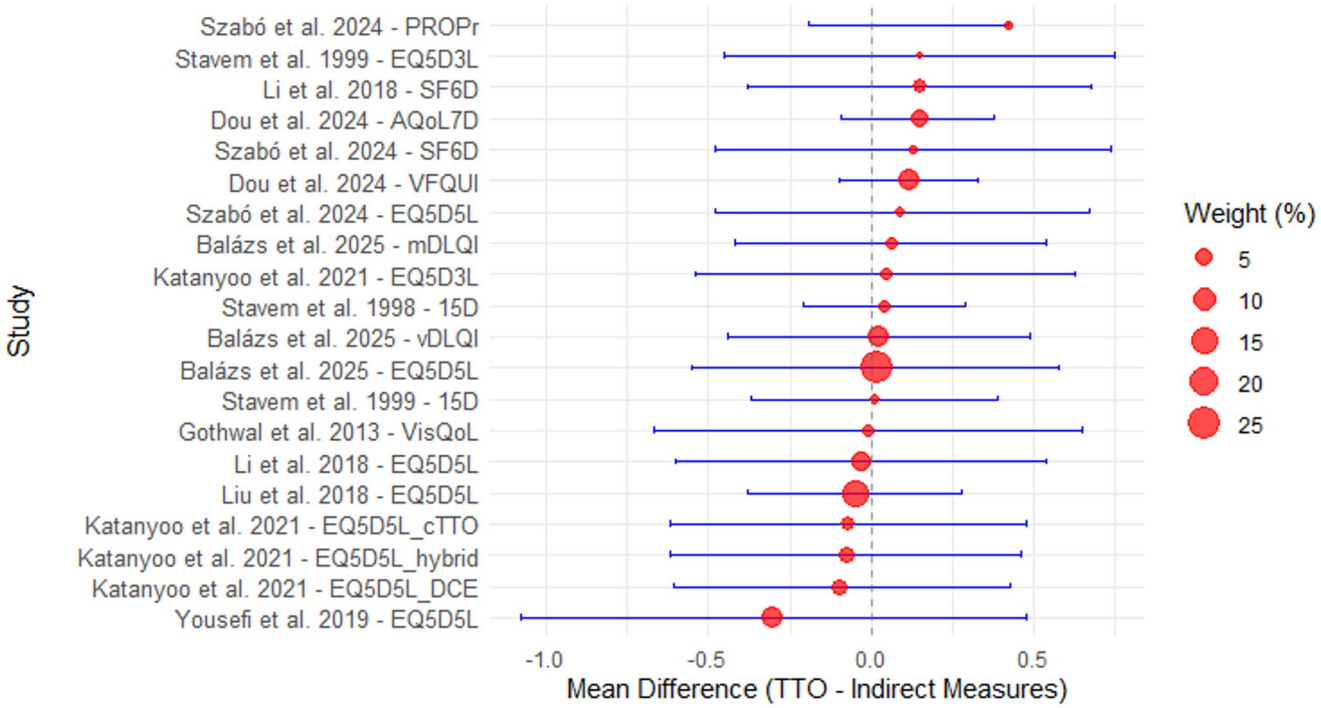

Fig. 3 Forest plot of overall mean difference between TTO and indirect utility measures. *The pooled mean difference was essentially zero, but study heterogeneity is (SD = 0.13, 95%CI = 0.09–0.20) pointing on considerably varying utility results study-to-study

Table 3 Measurement agreement parameters and utility values

Study	TTO mean (SD)	Measurement tool(s)*	Indirect measurement means (SD)	Indirect utility weight (tariff)	B-A mean difference (LOA)	Correlations: ICC (95%CoI) or Rho (p-level)	Conclusion on measurement agreement
Balázs et al. 2025	0.83 (0.24)	EQ-5D-5L	0.81 (0.24)	Hungarian	0.016 (−0.55–0.58)	ICC = 0.45 (0.36–0.52)	Agreement only between TTO and EQ-5D-5L (TTO advised in dermatology)
		mapping-DLQI	0.77 (0.13)	United Kingdom	0.063 (−0.42–0.54)	ICC = 0.31 (0.21–0.41)	
		value set-DLQI	0.81 (0.08)	Hungarian	0.025 (−0.44–0.49)	ICC = 0.26 (0.15–0.36)	
Dou et al. 2024	0.95 (0.06)	AQoL-7D	0.80 (0.11)	Austral	0.150 (−0.09–0.38)	ICC = 0.14 (nr.)	No agreement between the four measures
		VFQ-UI	0.83 (0.10)	International	0.120 (−0.10–0.33)	ICC = 0.16 (nr.)	
		SG	0.95 (0.10)	-	−0.001 (−0.21–0.21)	ICC = 0.11 (nr.)	
Szabó et al. 2024	0.89 (0.23)	EQ-5D-5L	0.79 (0.25)	United States	0.092 (−0.48–0.67)	ICC = 0.24 (0.07–0.40)	Poor agreement with all indirect measures (EQ-5D is advised)
		SF-6D	0.76 (0.21)	United States	0.130 (−0.48–0.74)	ICC = 0.20 (0.03–0.36)	
		PROPr	0.47 (0.24)	United States	0.420 (−0.19–0.42)	ICC = 0.06 (0.05–0.19)	
Katanyoo et al. 2021	0.80 (0.28)	EQ-5D-3L	0.76 (0.25)	Thai	0.046 (−0.54–0.63)	ICC = 0.53 (0.37–0.64)	Poor/inconclusive agreement stated with EQ-5D and VAS
		EQ-5D-5L(cTTO)	0.87 (0.18)		−0.070 (−0.62–0.48)	ICC = 0.45 (0.28–0.59)	
		EQ-5D-5L(DCE)	0.90 (0.14)		−0.099 (−0.62–0.43)	ICC = 0.40 (0.20–0.55)	
		EQ-5D-5L(hybrid)	0.88 (0.17)		−0.074 (−0.61–0.46)	ICC = 0.45 (0.28–0.59)	
		VAS	0.81 (0.16)	-	−0.005 (−0.52–0.51)	ICC = 0.52 (0.36–0.64)	
Yousefi et al. 2019	0.42 (0.47)	EQ-5D-5L	0.72 (0.13)	Iranian (crosswalk)	−0.301 (−1.08–0.48)	ICC = 0.50 (nr.)	No agreement
Li et al. 2018	0.80 (0.25)	EQ-5D-5L	0.83 (0.18)	Chinese	−0.030 (−0.60–0.54)	ICC = 0.25 (nr.)	No agreement (EQ-5D-5L is advised)
		SF-6D	0.65 (0.13)	Chinese	0.150 (−0.38–0.68)	ICC = 0.12 (nr.)	
Liu et al. 2018	0.85 (0.15)	EQ-5D-5L	0.90 (0.10)	Chinese	−0.050 (−0.38–0.28)	ICC = 0.27 (nr.)	No agreement (EQ-5D-5L is advised)
		SG	0.88 (0.20)	-	−0.030 (−0.50–0.44)	ICC = 0.24 (nr.)	
Kim et al. 2017	0.50 (0.35)	SG	0.54 (0.29)	-	−0.034 (−0.14–0.07)	ICC = 0.87 (nr.)	Agreement shown (SG is advised)
Li et al. 2014	0.96 (0.05)	SG	0.93 (0.09)	-	0.030 (−0.15–0.21)	NA	No agreement
Gothwal et al. 2013	0.65 (0.31)	VisQoL	0.66 (0.26)	Australia	−0.008 (−0.67–0.65)	NA	No agreement (VisQol is advised)
Stavem et al. 1998	0.92 (0.11)	15D	0.89 (0.09)	Finnish	0.040 (−0.21–0.29)	rho = 0.19 ($p > 0.05$)	Agreement between TTO and SG
		SG	0.93 (0.11)	-	−0.010 (−0.25–0.22)	rho = 0.48 ($p > 0.05$)	
		VAS	0.74 (0.17)	-	0.180 (−0.19–0.55)	rho = 0.19 ($p > 0.05$)	
Stavem et al. 1999	0.80 (nr.)	EQ-5D-3L	0.65 (nr.)	United Kingdom	0.150 (−0.45–0.75)	rho = 0.24 ($p > 0.05$)	Poor agreement of TTO with EQ-5D and SG
		15D	0.79 (nr.)	Finnish	0.010 (−0.37–0.39)	rho = 0.33 ($p < 0.01$)	
		SG	0.86 (nr.)	-	−0.060 (−0.30–0.18)	rho = 0.60 ($p < 0.01$)	

*AQoL-7d = Assessment of Quality-of-Life questionnaire, cTTO = composite time trade-off task, DCE = discrete choice experiment, DLQI = Dermatology Life Quality Index, SF-6D = Short Form 6 Dimension questionnaire, SG = standard gamble, TTO = time trade-off, VAS = visual analogue scale, VFQ-UI = Visual Function Questionnaire–Utility Index, VisQol = Vision-related Quality of Life Index

** CoI = confidence interval, NA = not available, LOA = limits of agreement, ICC = intraclass correlation coefficient

$N = 12$ studies, mean difference among TTO and direct/indirect utility measurements centres around zero but with wide 95%CI crossing 0, thus the absence of systemic difference is unclear. Albeit, on average TTO and other direct/indirect utility measures yield similar results, the between study heterogeneity suggest that individual study utility results vary slightly between TTO and direct (SD = ± 0.04) and greatly (by SD = ± 0.13) between TTO and indirect utility measures. Despite the small pooled mean differences, the considerable heterogeneity implies that individual studies may output very different utilities, large enough (in the range of ± 0.1–0.2) to influence QALY estimates and consequently cost-effectiveness conclusions. Our results support previous evidence, that agreement between TTO and direct/indirect utility measures is not consistent [4]. We did not find covariates that impact mean difference by populations, diseases, TTO method and data collection, perhaps only due to the low number of observations in the reference groups, thus a more focused investigation is required. Investigations of mean difference sources might start with TTO method

Table 4 Study quality rating

Study	JBI-1: inclusion criteria*	JBI-2: study setting detailed	JBI-4: measurement appropriateness	JBI-7: outcomes validity and reliability	JBI-8: adequate statistical analysis	rating score
Balázs et al. 2025	1	1	1	1	1	***5***
Dou et al. 2024	1	1	0	1	1	***4***
Gothwal et al. 2013	1	1	1	1	1	***5***
Katanyoo et al. 2021	1	1	1	1	1	***5***
Kim et al. 2017	1	1	1	1	1	***5***
Li et al. 2014	1	1	0	1	1	***4***
Li et al. 2018	1	1	1	1	1	***5***
Liu et al. 2018	1	1	1	1	1	***5***
Stavem et al. 1998	1	1	1	1	1	***5***
Stavem et al. 1999	1	1	1	1	1	***5***
Szabó et al. 2024	1	1	1	1	1	***5***
Yousefi et al. 2019	1	1	1	1	1	***5***

*JBI = Joanna Briggs Institute (Checklist for Analytical Cross Sectional Studies)

Table 5 Bayesian meta-analysis results of TTO and direct/indirect utility measures means differences

TTO—direct measures meta-analysis	**Estimated parameter**	**Overall mean difference**	SE	**Lower–Upper 95% credible intervals**	**Model fit: Bulk and Tail values**
	overall mean difference	−0.01	0.02	−0.04–0.03	1620; 2080
	between-study heterogeneity	0.04	0.02	0.02–0.08	1580; 2100
TTO-direct measures meta-regression	*Predictors*	*Estimated mean difference*	*SE*	*lower–upper 95% credible intervals*	*Regression fit: Bulk and Tail values*
	data collection (1 = interview)	−0.01	0.07	−0.14–0.13	5770; 5060
	type of TTO (1 = conventional)	0.01	0.08	−0.15–0.17	7070; 5320
	population (1 = patient)	0.01	0.08	−0.14–0.17	6610; 5150
	anchor state (1 = full health)	−0.01	0.07	−0.14–0.12	5060; 5180
	disease (1 = cancer)	−0.01	0.06	−0.11–0.11	5030; 4480
	intercept	−0.02	0.13	−0.27–0.23	5760; 5180
TTO-indirect measures meta-analysis	**Estimated parameter**	**Overall mean difference**	SE	**Lower–Upper 95% credible intervals**	**Model fit: Bulk and Tail values**
	utility mean difference	< 0.01	0.04	−0.08–0.09	1240; 1730
	between-study heterogeneity	0.13	0.03	0.09–0.20	1680; 3080
TTO-indirect measures meta-regression	*Predictors*	*Estimated mean difference*	*SE*	*lower–upper 95% credible intervals*	*Regression fit: Bulk and Tail values*
	data collection (1 = interview)	−0.07	0.07	−0.21–0.08	6760; 5320
	type of TTO (1 = conventional)	0.12	0.08	−0.04–0.28	7490; 5610
	population (1 = patient)	−0.09	0.07	−0.23–0.06	5770; 5380
	anchor state (1 = full health)	−0.06	0.07	−0.19–0.08	6870; 5590
	disease (1 = cancer)	−0.06	0.05	−0.17–0.05	6000; 5270
	intercept	0.06	0.14	−0.21–0.33	7650; 5380

attributes, but HRQoL instruments descriptive or rating systems and different value set/mapping algorithms might also skew the utility scores [80].

The large variety of TTO methodological attributes customized by the researchers, reinforce the major problem of incomparability of TTO utility values across studies. More studies prove that altered TTO task design has influence on respondents' health valuation. While the former Measurement and Valuation of Health (MVH) and the currently in use EuroQol Valuation Technology (EQ-VT) protocol

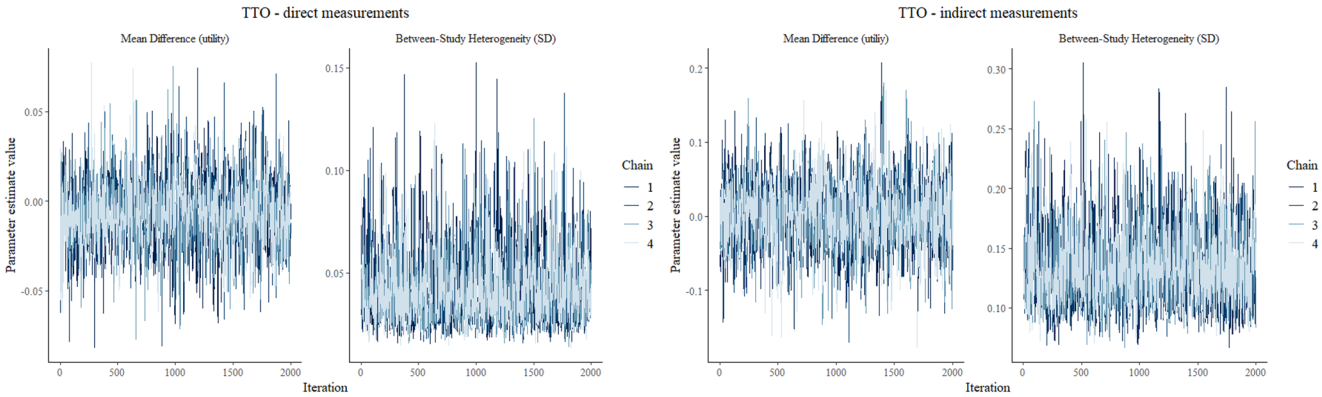

Fig. 4 MCMC trace plots of the Bayesian meta-analysis model. *The plots show excellent model convergance, the (posterior) overall mean difference estimates, forest plot and credible interval results are stable and reliable

describes a 'gold standard' TTO task, for valuation of EQ-5D health states, the recommendation is not adopted generally to TTO task design [81]. A methodological standardization would be critical to ensure comparability of studies health state assessments and prevent further exploration towards a best fitting set of TTO attributes [82]. Among the twelve included studies for comparison, only half complies with EQ-VT or MVH standards for TTO framework [67, 68, 70, 74, 77, 78].

This study is limited in not taking into consideration the study-specific weighting of indirect utility scores, although some used non-country-specific value-sets. The second limitation stands the scope of the review, as only 12 studies met the eligibility criteria, thus the small evidence has constraints in generalizability of findings. Only studies applying Bland–Altman agreement analysis, were deemed eligible as it is difficult to segment measurement agreement definition (often including comparison of construct-convergent validity correlation scores), moreover alternative qualitative agreement studies were disregarded, narrowing the included evidence. The generalizability of the meta-analysis is limited by the heterogeneity of TTO tasks across included studies, despite the meta-regression analysis, residual between-study variability may not be fully captured.

Between TTO and other direct/indirect utility measures our review found small mean differences, but significant between-study heterogeneity, indicating that measurement agreement is not consistent and utility means can vary substantially in both directions. Potential further research may explore the underlying reasons for mean differences, including methodological features of the TTO task, choice of value sets, and HRQoL instrument properties. Currently, whether discrepancies arise from valuation technique, instrument properties, or study context (and to what extent) remained undiscovered.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10198-026-01928-1.

Acknowledgements We are grateful for the help of Viktória Válay in the dataset management.

Author contributions P.B. contributed to perform the advanced search, and did the screening process separately with V.B. Data extraction and drafting of the first version of the manuscript was done by P.B while V.B. supervised the research, along with tables and figures design. No artificial intelligence (or AI-assisted technologies) was used in the production of the work.

Funding Open access funding provided by Corvinus University of Budapest. This research did not receive any specific funds.

Data availability Ensuring transparency all extracted data and R code is available in the supplementary material.

Declarations

Ethic statement Ethical approval was not sought for the present review study, as only publicly available information was analysed, without patient level data.

Competing interests The authors declare that they have no known conflict of financial or personal interests that might influence the research paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Brazier, J.: Valuing Health States for Use in Cost-Effectiveness Analysis. PharmacoEconomics (2008). https://doi.org/10.2165/00019053-200826090-00007
2. Dolan, P.: The measurement of individual utility and social welfare. J. Health Econ. (1998). https://doi.org/10.1016/S0167-6296(97)00022-2
3. Bansback, N., Brazier, J., Tsuchiya, A., Anis, A.: Using a discrete choice experiment to estimate health state utility values. J. Health Econ. (2012). https://doi.org/10.1016/j.jhealeco.2011.11.004
4. Arnold, D., Girling, A., Stevens, A., Lilford, R.: Comparison of direct and indirect methods of estimating health state utilities for resource allocation: review and empirical analysis. BMJ (2009). https://doi.org/10.1136/bmj.b2688
5. Davis, J.C., Liu-Ambrose, T., Khan, K.M., Robertson, M.C., Marra, C.A.: SF-6D and EQ-5D result in widely divergent incremental cost-effectiveness ratios in a clinical trial of older women: implications for health policy decisions. Osteoporos Int. (2012). https://doi.org/10.1007/s00198-011-1770-3
6. Torrance, G.W.: Measurement of health state utilities for economic appraisal: A review. J. Health Econ. (1986). https://doi.org/10.1016/0167-6296(86)90020-2
7. Attema, A.E., Versteegh, M.M., Oppe, M., Brouwer, W.B., Stolk, E.A.: Lead time TTO: leading to better health state valuations? Health Econ. (2013). https://doi.org/10.1002/hec.2804
8. Attema, A.E., Edelaar-Peeters, Y., Versteegh, M.M., Stolk, E.A.: Time trade-off: one methodology, different methods. Eur. J. Health Econ. (2013). https://doi.org/10.1007/s10198-013-0508-x
9. Alex G, M.J.,Wyrwich, K.W.: Health utility measures and the standard gamble. Academic emergency medicine (2003). https://doi.org/10.1111/j.1553-2712.2003.tb01349.x
10. Åström, M., ThetLwin, Z.M., Teni, F.S., Burström, K., Berg, J.: Use of the visual analogue scale for health state valuation: a scoping review. Qual. Life Res. (2023). https://doi.org/10.1007/s11136-023-03411-3
11. EuroQol Group: EuroQol-a new facility for the measurement of health-related quality of life. Health Policy (1990). https://doi.org/10.1016/0168-8510(90)90421-9
12. Brazier, J., Roberts, J., Deverill, M.: The estimation of a preference-based measure of health from the SF-36. J. Health Econ. (2002). https://doi.org/10.1016/s0167-6296(01)00130-8
13. Burström, K., Johannesson, M., Diderichsen, F.: A comparison of individual and social time trade-off values for health states in the general population. Health Policy (2006). https://doi.org/10.1016/j.healthpol.2005.06.011
14. Crott, R., Briggs, A.: Mapping the QLQ-C30 quality of life cancer questionnaire to EQ-5D patient preferences. Eur. J. Health Econ. (2010). https://doi.org/10.1007/s10198-010-0233-7
15. Boye, K.S., Matza, L.S., Feeny, D.H., Johnston, J.A., Bowman, L., Jordan, J.B.: Challenges to time trade-off utility assessment methods: when should you consider alternative approaches? Expert Rev. Pharmacoecon. Outcomes Res. (2014). https://doi.org/10.1586/14737167.2014.912562
16. Whitehurst, D.G., Bryan, S.: Another study showing that two preference-based measures of health-related quality of life (EQ-5D and SF-6D) are not interchangeable. But why should we expect them to be? Value Health (2011). https://doi.org/10.1016/j.jval.2010.09.002
17. Grieve, R., Grishchenko, M., Cairns, J.: SF-6D versus EQ-5D: reasons for differences in utility scores and impact on reported cost-utility. Eur. J. Health Econ. (2009). https://doi.org/10.1007/s10198-008-0097-2
18. Mulhern, B., Feng, Y., Shah, K., Janssen, M.F., Herdman, M., van Hout, B., et al.: Comparing the UK EQ-5D-3L and English EQ-5D-5L value sets. Pharmacoeconomics (2018). https://doi.org/10.1007/s40273-018-0628-3
19. Bland, J.M., Altman, D.: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet (1986). https://doi.org/10.1016/S0140-6736(86)90837-8
20. Zaki, R., Bulgiba, A., Ismail, R., Ismail, N.A.: Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: a systematic review. PLoS ONE (2012). https://doi.org/10.1371/journal.pone.0037908
21. Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. bmj (2021). https://doi.org/10.1136/bmj.n71
22. Mokkink, L.B., Terwee, C.B., Patrick, D.L., Alonso, J., Stratford, P.W., Knol, D.L., et al.: The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual. Life Res. (2010). https://doi.org/10.1007/s11136-010-9606-8
23. Bland, J.M., Altman, D.G.: Statistical methods for assessing agreement between two methods of clinical measurement. Int. J. Nurs. Stud. (2010). https://doi.org/10.1016/j.ijnurstu.2009.10.001
24. Oppe, M., Devlin, N.J., van Hout, B., Krabbe, P.F., de Charro, F.: A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health (2014). https://doi.org/10.1016/j.jval.2014.04.002
25. Ferreira, L.N., Ferreira, P.L., Pereira, L.N., Rowen, D., Brazier, J.E.: Exploring the consistency of the SF-6D. Value Health (2013). https://doi.org/10.1016/j.jval.2013.06.018
26. Sintonen, H.: The 15D instrument of health-related quality of life: properties and applications. Ann. Med. (2001). https://doi.org/10.3109/07853890109002086
27. Dewitt, B., Feeny, D., Fischhoff, B., Cella, D., Hays, R.D., Hess, R., et al.: Estimation of a preference-based summary score for the patient-reported outcomes measurement information system: the PROMIS®-preference (PROPr) scoring system. Med. Decis. Making (2018). https://doi.org/10.1177/0272989x18776637
28. Richardson, J., Angelo, I., Stuart, P., Kompal, S., Munir, K., Rose-Anne, M., et al.: Utility weights for the vision-related assessment of quality of life (AQoL)-7D instrument. Ophthalmic Epidemiol. (2012). https://doi.org/10.3109/09286586.2012.674613
29. Misajon, R., Hawthorne, G., Richardson, J., Barton, J., Peacock, S., Iezzi, A., et al.: Vision and quality of life: the development of a utility measure. Invest. Ophthalmol. Vis. Sci. (2005). https://doi.org/10.1167/iovs.04-1389
30. Peacock, S., RoseAnne, M., Angelo, I., Jeff, R., Graeme, H., Keeffe, J.: Vision and quality of life: development of methods for the VisQoL vision-related utility instrument. Ophthalmic Epidemiol. (2008). https://doi.org/10.1080/09286580801979417
31. Rentz, A.M., Kowalski, J.W., Walt, J.G., Hays, R.D., Brazier, J.E., Yu, R., et al.: Development of a preference-based index from the National Eye Institute Visual Function Questionnaire–25. JAMA Ophthalmol. (2014). https://doi.org/10.1001/jamaophthalmol.2013.7639
32. Finlay, A.Y., Khan, G.K.: Dermatology life quality index (DLQI)—a simple practical measure for routine clinical use. Clin. Exp. Dermatol. (1994). https://doi.org/10.1111/j.1365-2230.1994.tb01167.x
33. Barker, T.H., Hasanoff, S., Aromataris, E., Stone, J.C., Leonardi-Bee, J., Sears, K., et al.: The revised JBI critical appraisal tool for the assessment of risk of bias for analytical cross-sectional studies. JBI Evidence Synthesis (2025). https://doi.org/10.11124/jbies-24-00523
34. Santos, J.A.R., Grant, R., Di Tanna, G.L.: Bayesian meta-analysis of health state utility values: a tutorial with a practical application

in heart failure. Pharmacoeconomics (2024). https://doi.org/10.1007/s40273-024-01387-7
35. Bürkner, P.-C.: brms: An R package for Bayesian multilevel models using Stan. Journal of statistical software (2017). https://doi.org/10.18637/jss.v080.i01
36. Augestad, L.A., Rand, K., Luo, N., Barra, M.: Using the choice sequence in time trade-off as discrete choices: do the two stories match? Value Health (2020). https://doi.org/10.1016/j.jval.2019.10.003
37. Crump, R.T., Lau, R., Cox, E., Currie, G., Panepinto, J.: Testing the feasibility of eliciting preferences for health states from adolescents using direct methods. BMC Pediatr. (2018). https://doi.org/10.1186/s12887-018-1179-7
38. Cuervo, J., Castejón, N., Khalaf, K.M., Waweru, C., Globe, D., Patrick, D.L.: Development of the incontinence utility index: estimating population-based utilities associated with urinary problems from the incontinence quality of life questionnaire and neurogenic module. Health Qual. Life Outcomes (2014). https://doi.org/10.1186/s12955-014-0147-7
39. Cunningham, S.J., Hunt, N.P.: A comparison of health state utilities for dentofacial deformity as derived from patients and members of the general public. Eur. J. Orthod. (2000). https://doi.org/10.1093/ejo/22.3.335
40. Badia, X., Monserrat, S., Roset, M., Herdman, M.: Feasibility, validity and test–retest reliability of scaling methods for health states: the visual analogue scale and the time trade-off. Qual. Life Res. (1999). https://doi.org/10.1023/a:1008952423122
41. Hwang, H.-F., Chen, C.-Y., Lin, M.-R.: Patient-proxy agreement on the health-related quality of life one year after traumatic brain injury. Arch. Phys. Med. Rehabil. (2017). https://doi.org/10.1016/j.apmr.2017.05.013
42. Karimi, M., Brazier, J., Paisley, S.: Effect of reflection and deliberation on health state values: a mixed-methods study. Value Health (2019). https://doi.org/10.1016/j.jval.2019.07.013
43. Pullenayegum, E.M., Pickard, A.S., Xie, F.: Latent class models reveal poor agreement between discrete-choice and time tradeoff preferences. Med. Decis. Making (2019). https://doi.org/10.1177/0272989X19841592
44. Robinson, A., Spencer, A.E., Pinto-Prades, J.L., Covey, J.A.: Exploring differences between TTO and DCE in the valuation of health states. Med. Decis. Making (2017). https://doi.org/10.1177/0272989X16668343
45. Salomon, J.A.: Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. Popul. Health Metrics (2003). https://doi.org/10.1186/1478-7954-1-12
46. Seidler, A.L., Rethberg, C., Schmitt, J., Nienhaus, A., Seidler, A.: Health utilities for chronic low back pain. J. Occup. Med. Toxicol. (2017). https://doi.org/10.1186/s12995-017-0172-7
47. Spencer, A., Tomeny, E., Mujica-Mota, R.E., Robinson, A., Covey, J., Pinto-Prades, J.L.: Do time trade-off values fully capture attitudes that are relevant to health-related choices? Eur. J. Health Econ. (2019). https://doi.org/10.1007/s10198-018-1017-8
48. Xie, S., Wu, J., Chen, G.: Discrete choice experiment with duration versus time trade-off: a comparison of test–retest reliability of health utility elicitation approaches in SF-6Dv2 valuation. Qual. Life Res. (2022). https://doi.org/10.1007/s11136-022-03159-2
49. Xie, S., Wu, J., He, X., Chen, G., Brazier, J.E.: Do discrete choice experiments approaches perform better than time trade-off in eliciting health state utilities? Evidence from SF-6Dv2 in China. Value Health (2020). https://doi.org/10.1016/j.jval.2020.06.010
50. Laupacis, A., Wong, C., Churchill, D., Group, C.E.S.: The use of generic and specific quality-of-life measures in hemodialysis patients treated with erythropoietin. Control. Clin. Trials (1991). https://doi.org/10.1016/s0197-2456(05)80021-2
51. Ameri, H., Poder, T.G.: Assessing the direct impact of death on discrete choice experiment utilities. Appl. Health Econ. Health Policy (2025). https://doi.org/10.1007/s40258-024-00929-6
52. Ariza-Ariza, R., Hernández-Cruz, B., Carmona, L., Dolores Ruiz-Montesinos, M., Ballina, J., Navarro-Sarabia, F.: Assessing utility values in rheumatoid arthritis: a comparison between time trade-off and the EuroQol. Arthritis Care & Research: Official Journal of the American College of Rheumatology (2006). https://doi.org/10.1002/art.22226
53. Bijlenga, D., Birnie, E., Bonsel, G.J.: Feasibility, reliability, and validity of three health-state valuation methods using multiple-outcome vignettes on moderate-risk pregnancy at term. Value Health (2009). https://doi.org/10.1111/j.1524-4733.2009.00503.x
54. Bijlenga, D., Birnie, E., Mol, B.W.J., Bonsel, G.J.: Obstetrical outcome valuations by patients, professionals, and laypersons: differences within and between groups using three valuation methods. BMC Pregnancy Childbirth (2011). https://doi.org/10.1186/1471-2393-11-93
55. Craig, B.M., Busschbach, J.J., Salomon, J.A.: Modeling ranking, time trade-off, and visual analog scale values for EQ-5D health states: a review and comparison of methods. Med. Care (2009). https://doi.org/10.1097/MLR.0b013e31819432ba
56. Daly, E., Gray, A., Barlow, D., McPherson, K., Roche, M., Vessey, M.: Measuring the impact of menopausal symptoms on quality of life. BMJ (1993). https://doi.org/10.1136/bmj.307.6908.836
57. Gregor, J.C., McDonald, J.W., Klar, N., Wall, R., Atkinson, K., Lamba, B., et al.: An evaluation of utility measurement in Crohn's disease. Inflamm. Bowel Dis. (1997). https://doi.org/10.1002/ibd.3780030405
58. Honkalampi, T., Sintonen, H.: Do the 15D scores and time trade-off (TTO) values of hospital patients' own health agree? Int. J. Technol. Assess. Health Care (2010). https://doi.org/10.1017/S0266462309990869
59. Hornberger, J.C., Redelmeier, D.A., Petersen, J.: Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. J. Clin. Epidemiol. (1992). https://doi.org/10.1016/0895-4356(92)90099-9
60. Khanna, D., Furst, D.E., Wong, W.K., Tsevat, J., Clements, P.J., Park, G.S., et al.: Reliability, validity, and minimally important differences of the SF-6D in systemic sclerosis. Qual. Life Res. (2007). https://doi.org/10.1007/s11136-007-9207-3
61. Lin, M.-R., Hwang, H.-F., Chung, K.-P., Huang, C., Chen, C.-Y.: Rating scale, standard gamble, and time trade-off for people with traumatic spinal cord injuries. Phys. Ther. (2006). https://doi.org/10.1093/ptj/86.3.337
62. Schwarzinger, M., Stouthard, M.E., Burström, K., Nord, E.,nl, E.D.W.G.v.f.f.e.: Cross-national agreement on disability weights: the European Disability Weights Project. Population Health Metrics (2003). https://doi.org/10.1186/1478-7954-1-9
63. Shah, K.K., Ramos-Goñi, J.M., Kreimeier, S., Devlin, N.J.: An exploration of methods for obtaining 0= dead anchors for latent scale EQ-5D-Y values. Eur. J. Health Econ. (2020). https://doi.org/10.1007/s10198-020-01205-9
64. Wu, J., Xie, S., He, X., Chen, G., Bai, G., Feng, D., et al.: Valuation of SF-6Dv2 health states in China using time trade-off and discrete-choice experiment with a duration dimension. Pharmacoeconomics (2021). https://doi.org/10.1007/s40273-020-00997-1
65. Yanal, N., Al Massri, A.M., Hammad, E.A.: Validity, reliability, and feasibility of EQ-5D-3L, VAS, and time trade-off among Jordanians. J. Healthc. Qual. Res. (2025). https://doi.org/10.1016/j.jhqr.2024.10.001
66. Moore, A., Clarke, A., Danoff, D., Joseph, L., Belisle, P., Neville, C., et al.: Can health utility measures be used in lupus research? A comparative validation and reliability study of 4 utility indices. The Journal of rheumatology (1999). no DOI

67. Balázs, P.G., Gáspár, K., Gergely, H.L., Hajdú, K., Holló, P., Koszorú, K., et al.: Comparison of health-related quality of life in atopic dermatitis, hidradenitis suppurativa, pemphigus and psoriasis. Arch. Dermatol. Res. (2025). https://doi.org/10.1007/s00403-024-03786-4
68. Szabó, Á., Brodszky, V., Rencz, F.: Comparing EQ-5D-5L, PROPr, SF-6D and TTO utilities in patients with chronic skin diseases. Eur. J. Health Econ. (2025). https://doi.org/10.1007/s10198-024-01728-5
69. Dou, L., Xu, Y., Chen, G., Li, S.: Psychometric properties and comparison of four health utility approaches among myopia patients in China. Health Qual. Life Outcomes (2023). https://doi.org/10.1186/s12955-023-02150-w
70. Katanyoo, K., Thavorncharoensap, M., Chaikledkaew, U., Riewpaiboon, A.: A comparison of six approaches for measuring utility values among patients with locally advanced cervical cancer. Expert Rev. Pharmacoecon. Outcomes Res. (2022). https://doi.org/10.1080/14737167.2021.1906224
71. Yousefi, M., Safari, H., Akbari Sari, A., Raei, B., Ameri, H.: Assessing the performance of direct and indirect utility eliciting methods in patients with colorectal cancer: EQ-5D-5L versus C-TTO. Health Serv. Outcomes Res. Method. (2019). https://doi.org/10.1007/s10742-019-00204-5
72. Li, S., Wang, M., Liu, L., Chen, G.: Which approach is better in eliciting health state utilities from breast cancer patients? Evidence from mainland China. Eur. J. Cancer Care (2019). https://doi.org/10.1111/ecc.12965
73. Liu, L., Li, S., Zhao, Y., Zhang, J., Chen, G.: Health state utilities and subjective well-being among psoriasis vulgaris patients in mainland China. Qual. Life Res. (2018). https://doi.org/10.1007/s11136-018-1819-2
74. Kim, S.-H., Lee, S.-i., Jo, M.-W.: Feasibility, comparability, and reliability of the standard gamble compared with the rating scale and time trade-off techniques in Korean population. Qual. Life Res. (2017). https://doi.org/10.1007/s11136-017-1676-4
75. Li, S., Wang, G., Xu, Y., Gray, A., Chen, G.: Utility values among myopic patients in mainland China. Optom. Vis. Sci. (2014). https://doi.org/10.1097/opx.0000000000000299
76. Gothwal, V.K., Bagga, D.K.: Utility values in the visually impaired: comparing time-trade off and VisQoL. Optom. Vis. Sci. (2013). https://doi.org/10.1097/OPX.0b013e318291063a
77. Stavem, K.: Reliability, validity and responsiveness of two multiattribute utility measures in patients with chronic obstructive pulmonary disease. Qual. Life Res. (1999). https://doi.org/10.1023/a:1026475531996
78. Stavem, K.: Quality of life in epilepsy: comparison of four preference measures. Epilepsy Res. (1998). https://doi.org/10.1016/S0920-1211(97)00075-2
79. Whitehurst, D.G., Bryan, S., Lewis, M.: Systematic review and empirical comparison of contemporaneous EQ-5D and SF-6D group mean scores. Med. Decis. Making (2011). https://doi.org/10.1177/0272989x11421529
80. Mukuria, C., Rowen, D., Harnan, S., Rawdin, A., Wong, R., Ara, R., et al.: An updated systematic review of studies mapping (or cross-walking) measures of health-related quality of life to generic preference-based measures to generate utility values. Appl. Health Econ. Health Policy (2019). https://doi.org/10.1007/s40258-019-00467-6
81. Oppe, M., Rand-Hendriksen, K., Shah, K., Ramos-Goñi, J.M., Luo, N.: EuroQol Protocols for Time Trade-Off Valuation of Health Outcomes. PharmacoEconomics (2016). https://doi.org/10.1007/s40273-016-0404-1
82. Lugnér, A.K., Krabbe, P.F.M.: An overview of the time trade-off method: concept, foundation, and the evaluation of distorting factors in putting a value on health. Expert Rev. Pharmacoecon. Outcomes Res. (2020). https://doi.org/10.1080/14737167.2020.1779062

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.