ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

On the Optimization of Bayesian D-Efficient DCE Designs for the Estimation of QALY Tariffs That are Corrected for Nonlinear Time Preferences

Marcel F. Jonker, PhD,[1,2,3,*] Michiel C.J. Bliemer, PhD[4]

[1]Duke Clinical Research Institute, Duke University, Durham, NC, USA; [2]Erasmus Choice Modeling Centre, Erasmus University Rotterdam, Rotterdam, The Netherlands; [3]Erasmus School of Health Policy & Management, Erasmus University Rotterdam, Rotterdam, The Netherlands; [4]Institute of Transport and Logistics, University of Sydney Business School, NSW 2006, Australia

ABSTRACT

Objectives: This article explains how to optimize Bayesian D-efficient discrete choice experiment (DCE) designs for the estimation of quality-adjusted life year (QALY) tariffs that are unconfounded by respondents' time preferences.

Methods: The calculation of Bayesian D-errors is explained for DCE designs that allow for the disentanglement of respondents' time and health-state preferences. Time preferences are modelled via an exponential, hyperbolic, or power discount function and the performance of the proposed DCE designs is compared with that of several conventional DCE designs that do not take nonlinear time preferences into account.

Results: Based on the achieved D-error, asymptotic standard error, and estimated sample size to obtain statistically significant estimates of the discount rate parameters, the proposed designs outperform the conventional DCE designs.

Conclusions: We recommend that applied researchers use appropriately optimized DCE designs for the estimation of QALY tariffs that are corrected for time preferences. The TPC-QALY software package that accompanies this article makes the recommended designs easily accessible for health-state valuation researchers.

Keywords: discrete choice experiment, health state valuation, quality-adjusted life year, time preferences.

VALUE HEALTH. 2019; ■(■):■–■

Introduction

Time-preference-corrected quality-adjusted life year (QALY) tariffs avoid confounding between quality of life and time preferences without resulting in time-dependent QALY tariffs.[1] In nominal terms, each QALY still represents 1 year in perfect health, which conforms to the conventional QALY assumptions. However, when QALYs are compared across time, they can and should be discounted to properly reflect time preferences. The latter is already standard practice in health technology assessment applications, which means that time-preference-corrected QALY tariffs more closely align with health technology assessment than traditional QALY tariffs that are derived under the assumption of linear time preferences.

Time-preference-corrected QALY tariffs are also preferable from a theoretical perspective. That is, linear time preferences are hardly ever observed in human decision making and are unrealistic to presume from the outset.[2] Moreover, empirical evidence seems to suggest that the assumption of linear time preferences does not hold in traditional time trade-off (TTO) or, particularly, in DCE-duration estimations.[1,3–5]

Incorrectly imposing linear time preferences does not appear to be very consequential when a TTO elicitation format is used. However, it can result in severely biased QALY tariffs when a DCE-duration elicitation format is used. Unlike with TTO, the fraction of health states that are valued as worse than immediate death is not directly observed in DCE-duration tasks. Instead, health states worse than immediate death are identified using a model-based extrapolation, which is sensitive to the assumptions made about respondents' time preferences. The latter explains why many of the QALY tariffs that have thus far been established with DCE-duration methods have a higher percentage of health states classified as worse than immediate death than occurs with TTO formats

The authors have indicated that they have no conflicts of interest with regard to the content of this article.

* Address correspondence to: Marcel F. Jonker, PhD, Duke Clinical Research Institute, 300 W. Morgan Street, Durham, NC 27701. Email: marcel@mfjonker.com

https://doi.org/10.1016/j.jval.2019.05.014

(see, eg, Jonker et al[6] and Mulhern et al[7] vs Versteegh et al[8] and Devlin et al[9]).

Time-preference-corrected QALY tariffs appear to be important to avoid downward biased QALY tariffs, particularly with DCE-duration elicitation formats. At the same time, an impediment to their implementation is that the estimation of time-preference-corrected QALY tariffs requires DCE designs that have adequate statistical efficiency to reliably elicit and disentangle respondents' time and health-state preferences. As shown in this article, standard DCE designs are inefficient for the estimation of time-preference-corrected QALY tariffs and can even result in identification problems. Accordingly, to avoid these problems from the outset, this article provides a complete exposition on how to create Bayesian D-efficient DCE designs that optimally accommodate the estimation of time-preference-corrected QALY tariffs. In addition, it provides an easy-to-use software implementation that includes several commonly used health-state valuation instruments, including the EQ-5D-5L[10] instrument.

Methods

Relevant Class of Utility Functions

Time-preference-corrected QALY tariffs[1] are derived from a class of health-state valuation utility functions that are defined as follows:

$$U_{ijt} = H_{ijt}.NPV_{ijt} + \varepsilon_{ijt}, \quad i = 1,..,I; j = 1,..,J; t = 1,..,T. \tag{1}$$

Here, the utility (U_{ijt}) that respondent i obtains from alternative j in choice task t is defined as the sum of a systematic component (H_{ijt} .NPV_{ijt}) and an unobserved error term (ε_{ijt}). The error term is assumed to be independently and identically extreme value type I distributed, and the systematic component describes how health-state utilities are derived from the multiplication of quality and quantity of life. More specifically:

1. The quality of life (ie, health-state values H_{ijt}) component is assumed to be modeled as a standard linear additive function that is defined as the dot product of K dummy coded health-state characteristics ($X_{ijt1},..,X_{ijtK}$) and associated vectors of coefficients ($\beta_1,..,\beta_K$):

$$H_{ijt} = \sum_{k=1}^{K} \beta_k.X_{ijtk}, \quad i = 1,..,I; j = 1,..,J; t = 1,..,T. \tag{2}$$

2. The quantity of life component is defined as the net present value (NPV_{ijt}) of the number of life years (Q_{ijt}) spent in each health state, which equals the sum of the present values (PV) of all future life years q=1,…,Q_{ijt}:

$$NPV_{ijt} = \sum_{q=1}^{Q_{ijt}} PV_q, \quad i = 1,..,I; j = 1,..,J; t = 1,..,T. \tag{3}$$

Note that Equation 3 is very general and implies that *any* discount function can be used to obtain the PV of future life years. The most commonly used discounting functions are the exponential,[11] hyperbolic,[12] and power[13] discounting functions (see Box 1). Each of

BOX 1. COMMONLY USED DISCOUNTING FUNCTIONS*

A. Exponential discounting:

$$NPV_{ijt} = \sum_{q=1}^{Q_{ijt}} PV_q(r) = \sum_{q=1}^{Q_{ijt}} exp(-r\cdot q) = \begin{cases} Q_{ijt} & , \ if\ r = 0 \\ \dfrac{1-\exp(-r\cdot Q_{ijt})}{\exp(r)-1} & , \ if\ r \neq 0 \end{cases}$$

$$\frac{\partial NPV_{ijt}}{\partial r} = \frac{Q_{ijt}\cdot exp(-r\cdot Q_{ijt})}{\exp(r)-1} - \frac{exp(r) * (1-\exp(-r\cdot Q_{ijt}))}{(\exp(r)-1)^2}, \ r \neq 0$$

B. Hyperbolic discounting[†,‡]:

$$NPV_{ijt} = \sum_{q=1}^{Q_{ijt}} PV_q(r) = \sum_{q=1}^{Q_{ijt}} \frac{1}{1+r\cdot q} = \begin{cases} Q_{ijt} & , \ if\ r = 0 \\ \dfrac{\psi\left(1+\frac{1}{r}+Q_{ijt}\right) - \psi\left(1+\frac{1}{r}\right)}{r} & , \ if\ r > 0 \end{cases}$$

$$\frac{\partial NPV_{ijt}}{\partial r} = \frac{-r\cdot\psi\left(1+\frac{1}{r}+Q_{ijt}\right) - \psi^1\left(1+\frac{1}{r}+Q_{ijt}\right) + r\cdot\psi\left(1+\frac{1}{r}\right) + \psi^1\left(1+\frac{1}{r}\right)}{r^3}, \ r > 0$$

C. Power discounting:

$$NPV_{ijt} = \sum_{q=1}^{Q_{ijt}} PV_q(r) = \sum_{q=1}^{Q_{ijt}} (q+1)^{1-r} - q^{1-r} = \begin{cases} Q_{ijt} & , if\ r = 0 \\ (Q_{ijt}+1)^{1-r} - 1 & , if\ r \neq 0 \end{cases}$$

$$\frac{\partial NPV_{ijt}}{\partial r} = (Q_{ijt}+1)^{1-r}\cdot\log(Q_{ijt}+1), \ r \neq 0$$

* ψ and ψ^1 denote the di- and tri-gamma function, respectively.
† For hyperbolic discount rates smaller than 0, the NPV needs to be calculated recursively.
‡ Partial derivatives with respect to the discount rate (r) are required for the calculation of the D-error.

BOX 2. BAYESIAN D-ERROR FOR A STANDARD CONDITIONAL LOGIT MODEL WITH A LINEAR ADDITIVE UTILITY FUNCTION

Utility function:	$U_{ijt} = \sum_{k=1}^{K} \beta_k \cdot X_{ijtk} + \varepsilon_{ijt},\ \varepsilon_{ijt} \sim EV1(0,1)$	(I)
Probability:	$P_{ijt} = \frac{\exp(U_{ijt})}{\sum_{a=1}^{J} \exp(U_{iat})}$	(II)
Partial derivatives:	$\frac{\partial U_{ijt}}{\partial \beta} = X_{ijt}$	(III)
Intermediate matrix:	$Z_{ijt} = (X_{ijt} - \sum_{a=1}^{J} (X_{iat} \cdot P_{iat})) \sqrt{P_{ijt}}$	(IV)
Bayesian D-error*:	$D\text{-}error = \int_{\beta} \det(Z'Z)^{(-1/K)} \varphi(\beta \mid \theta) d\beta$	(V)

* In the Bayesian D-error calculations, the β parameters are random variables with a joint probability density function φ (.) with given parameters θ. Usually, the β parameters are assumed to be multivariate normal distributed, ie, $\beta \sim$ MVN(μ,Σ).

these has 1 input parameter, the discount rate (r), which controls the degree of discounting. The higher the discount rate, the less weight is attached to life years spent in the more distant future. In contrast, when all future life years are valued equally, the discount rate will be 0 and the NPV will be equal to Q. Accordingly, the traditional QALY assumption of perfectly linear time preferences is embedded as a special case.

The Bayesian D-Error Optimization Criterion

The objective of this article is to present an experimental design with attribute values (X_{ijt1},...,X_{ijtK} and Q_{ijt}) for each respondent i, alternative j, and choice task t that allows for the efficient estimation of the health-state preference parameters (β) and discount rate (r). Several different efficiency measures have been proposed in the literature, but the most widely used efficiency measure is the D-error. The D-error takes the determinant of the asymptotic variance-covariance matrix and scales it with the number of parameters to be estimated. In this article, we focus solely on the D-error, although alternative efficiency criteria could also be used; see Scarpa and Rose[14] for a comparison of alternative design criteria.

In contrast to efficient DCE designs for standard linear additive utility functions, for which the appropriate D-error calculations are described in multiple publications (eg, Scarpa and Rose,[14] Kanninen,[15] and Rose and Bliemer[16]) and included in several software packages (eg, ChoiceMetrics [Ngene software, Australia] and JMP [SAS Institute Inc, Cary, NC]), there is currently no algorithm or software available that can optimize for the nonlinear multiplicative utility functions as described by Equations 1 to 3. The reason is that a linear additive utility function greatly simplifies the D-efficiency calculations, which, vise versa, implies that more complex D-error calculations are required for nonlinear multiplicative utility functions.

More specifically, for a standard conditional logit model with a linear additive utility function ($U_{ijt} = \sum_{k=1}^{K} \beta_k . X_{ijtk}$), the matrix of partial derivatives with respect to the estimated model parameters (ie, β) equals X. This property is used in the calculation of the D-error for the conditional logit model as described in Box 2. However, the matrix of partial derivatives of the utility function as described by Equations 1 to 3 with respect to the parameters to be estimated (ie, β and r) does not simplify to X. Hence, the standard D-error calculations in Box 2 cannot be used for the optimization of efficient DCE designs for time-preference-corrected QALY tariffs; instead, the calculations as described in Box 3 need to be used.

As shown in Box 3, the calculation of the D-efficiency for the nonlinear multiplicative utility function requires the derivation of the matrix of partial derivatives $\partial U_{ijt}/\partial \gamma$ for all parameters to be estimated (ie, β and r). This matrix does not reduce to X and depends on the duration and type of discount function used. Accordingly, a D-efficient DCE design that is optimized for 1 type of discount function is not necessarily efficient for other discounting functions. More important, D-efficient designs that are optimized assuming a linear additive utility function (eg, using currently available software packages) will not take the

BOX 3. BAYESIAN D-ERROR FOR A CONDITIONAL LOGIT MODEL WITH A NONLINEAR MULTIPLICATIVE UTILITY FUNCTION

Utility function:	$U_{ijt} = (\sum_{k=1}^{K} \beta_k \cdot X_{ijtk}) \cdot NPV_{ijt} + \varepsilon_{ijt},\ \varepsilon_{ijt} \sim EV1(0,1)$	(I)
Probability:	$P_{ijt} = \frac{\exp(U_{ijt})}{\sum_{a=1}^{J} \exp(U_{iat})}$	(II)
Parameter vector:	$\gamma = (\beta \cup r)$	(III)
Intermediate matrix:	$Z_{ijt} = \left(\frac{\partial U_{ijt}}{\partial \gamma} - \sum_{a=1}^{J} \left(\frac{\partial U_{iat}}{\partial \gamma} \cdot P_{iat} \right) \right) \sqrt{P_{ijt}}$	(IV)
Bayesian D-error*:	$D\text{-}error = \int_{\gamma} \det(Z'Z)^{(-1/(K+1))} \varphi(\gamma \mid \theta) d\gamma$	(V)

* In the Bayesian D-error calculations, the γ parameters are random variables with a joint probability density function φ (.) with given parameters θ. Here, all γ parameters are assumed to be multivariate normal distributed, ie, $\gamma \sim$ MVN(μ,Σ).

appropriate derivates into account and are unlikely to be D-efficient for *any* multiplicative utility function.

Additional DCE Design Considerations

In addition to the appropriate design optimization criterion, several other considerations are important for the optimization of DCE designs for the estimation of both standard and time-preference-corrected QALY tariffs:

1. Even when a DCE duration design is optimized for a multiplicative utility function, unconstrained elicitation formats provide no guarantee that respondents do not adopt a simpler, linear additive utility function instead of the required multiplicative utility function for the QALY calculations. A linear additive utility function would reduce the task complexity for respondents but would theoretically invalidate QALY tariff calculations based on the observed choice data. To preemptively avoid this problem, Jonker et al[1,6] used a "matched-pairs" choice format with 2 types of choice tasks: (1) pairwise choice tasks consisting of different impaired health states with an equal duration of life, and (2) pairwise choice tasks in which an impaired health state with longer duration is compared with perfect health in a shorter duration. In the matched-pairs format, the 2 types of formats are linked by the imposition that the impaired health state in the second choice task is identical to one of the impaired health states in the first choice task and presented in a single layout. This simplifies the second choice task for respondents, but the essential point is that the DCE design contains sufficient overlap in the duration levels to ensure that respondents make choices that adhere (at least approximately) to the required multiplicative utility function.
2. Furthermore, even with sufficient overlap in duration, it is advisable to impose some overlap in the health-state attributes as well. Although the introduction of level overlap reduces the statistical efficiency of the DCE design, it also reduces the complexity of the choice tasks for respondents and consequently improves behavioral efficiency—that is, it tends to reduce the drop-out rate, increase the level of choice consistency, and avoids problems with attribute nonattendance.[17,18] These improvements in behavioral efficiency mitigate the loss in statistical efficiency and increase the quality of the collected choice data.
3. Another important consideration is the benefit of a severity-stratified selection of health states in the DCE design. Unlike with TTO, where the position of immediate death can be directly observed, the position of immediate death is based on an extrapolation when using DCE duration. The use of a severity-stratified DCE design can improve the robustness of the DCE design and avoid biased estimates when the utility function is misspecified.[19] Hence, some type of severity stratification is advisable, particularly when (a priori) the correct model specification and/or discount function is unknown.
4. Finally, many DCE health-state valuation studies have used a single DCE design, which is shown in its entirety to all participating respondents. Sándor and Wedel[20] have shown that it is more efficient and robust to simultaneously optimize multiple versions of a DCE design and assign each respondent (randomly) to only one of the versions. These so-called heterogeneous DCE designs differ from traditional blocked designs in the sense that each subdesign is a stand-alone design as opposed to merely being a fraction of a stand-alone design. The advantage of heterogeneous compared with homogeneous DCE designs is that heterogeneous designs increase statistical efficiency without increasing the survey burden for participating respondents; each individual respondent completes only one of the subdesigns.

Fortran Implementation

The optimization of Bayesian D-efficient DCE designs for the estimation of time-preference-corrected QALY tariffs has been implemented in Fortran, with built-in support for several health-state instruments and 3 different nonlinear discount functions (ie, exponential, hyperbolic, and power). The previously described design considerations are implemented as follows:

1. Overlap in duration is included by optimizing for the matched-pairs format.
2. The algorithm supports a flexible amount of attribute-level overlap.
3. Based on the supplied priors, the algorithm automatically implements a severity-stratified selection of health states (cf Lim et al[19]).
4. The algorithm creates heterogeneous DCE designs using an optimization criterion that is a weighted average of the overall D-error and the average D-error of the individual subdesigns.

Furthermore, the design optimization can be specified with various numbers of quasi-random draws from the specified priors to evaluate the Bayesian D-efficiency criterion, and it automatically generates an optimized Latin hypercube sample that has good sampling properties.[21] To avoid left–right bias in the health-state selection, the design criterion includes comparisons between options A and B, B and C, and A and C (ie, a full ranking of the choice options in each choice task).

Generating a time-preference-corrected QALY design with the Fortran TPC-QALY software package requires the following steps. First, after opening the program, which requires a Windows operating system, the user has to select the type of instrument, type of discount function, number of matched pairs, and number of subdesigns. On the next screen, the user needs to specify the amount of level overlap, number of Bayesian draws, duration levels, and the weights to be used for the heterogeneous design optimization. The default and recommended setting is to use an optimization criterion that is for 75% based on the average D-error of the subdesigns and 25% on the D-error of overall design. Finally, the Bayesian priors need to be supplied. Note that the TPC-QALY program imposes a minimum amount of structure on the specified priors to avoid accidental misspecification (eg, by taking the ordinal structure of the attributes into account). After the user clicks on the "optimize" button, the design optimization will run for a few minutes up to several hours (depending on the computer speed and the number of Bayesian draws), after which the program saves the TPC-QALY DCE design as a .txt file that can be imported by standard survey software, such as, for example, Sawtooth Software (Sawtooth Software Inc, Provo, UT). (For a more detailed description of the TPC-QALY program, see Appendix C in the Supplementary Materials found at https://doi.org/10.1016/j.jval.2019.05.014).

Efficiency Comparisons

To establish the efficiency improvement of TPC-QALY versus standard DCE designs for the estimation of time-preference-corrected QALY tariffs, we compared a total of 8 DCE designs. All design comparisons were based on the largest and most commonly used instrument that is currently included in the TCP-QALY software package: the EQ-5D-5L instrument.[10] Four different designs were optimized using the Bayesian D-error criterion as

described in Box 3—that is, 3 designs assuming an exponential, hyperbolic, and power function and a fourth design that was optimized for all of these discount functions simultaneously. In addition, the following 4 conventional DCE designs were included:

1. A Bayesian D-efficient linear multiplicative DCE design with utility function

$$U_{ijt} = \left[\sum_{k=1}^{K} \beta_k . X_{ijtk}\right] . Q_{ijt} + \varepsilon_{ijt}, \quad i = 1, .., I; j = 1, .., J; t = 1, .., T. \tag{4}$$

that correctly incorporates the multiplication of quality and quantity of life while imposing linear time preferences. Designs optimized using this criterion have been used in, for example, Jonker et al[6] and Lim et al.[19]

2. Three different Bayesian D-efficient linear additive DCE designs with utility function

$$U_{ijt} = \sum_{k=1}^{K} \beta_k . X_{ijtk} + \beta_{K+1} . Q_{ijt} + \varepsilon_{ijt}, i = 1, .., I; j = 1, .., J; t = 1, .., T, \tag{5}$$

in which duration is included as an additional linear additive attribute; that is, a heterogeneous efficient design with severity stratification, a homogeneous efficient design without severity stratification, and a homogeneous efficient design without severity stratification that was optimized using 0 priors.

All DCE designs were optimized using the default settings of the TPC-QALY program: 100 Bayesian draws, 2 levels overlapped, 12 matched pairwise choice tasks, and, for the heterogeneous DCE designs, 10 subdesigns with 75% optimization weight on the average D-error of the individual subdesigns and 25% on the overall D-error of the DCE design. A standard modified Fedorov optimization algorithm was used to optimize the health states. The duration levels were fixed at (/2,3,5,7,8,9,10,11,12,12,15,15/) years for choice options A and B in the matched-pairs format, with the levels randomly distributed across choice tasks. The duration levels for option C were selected from all integers smaller than the corresponding duration levels of options A and B and additionally comprised 3 and 6 months of duration. A greedy optimization of the duration levels was performed by evaluating all possible duration levels every 10 000 optimization iterations.

The reported design efficiencies were based on the best achieved D-error after 3 separate runs of 200 000 optimization iterations each. The performance of the constructed DCE designs was subsequently evaluated in terms of the achieved D-error for the estimation of time-preference-corrected QALY tariffs and the achieved asymptotic standard error and estimated sample size for obtaining statistically significant estimates for the discount rate parameter. Accordingly, both the overall performance of the DCE designs and the designs' ability to disentangle respondents' time and health-state preferences were investigated.

The priors for the Bayesian efficient design optimizations are listed in Appendix A (see the Supplementary Materials found at https://doi.org/10.1016/j.jval.2019.05.014). These were obtained by fitting a conditional logit model with each of the included utility functions on the DCE-duration data that were previously used by Lim et al.[19] These data were collected using a severity-stratified heterogeneous DCE design with 8 subdesigns, which was optimized using the standard multiplicative utility function as described in Equation 4 with 21 matched pairwise choice tasks per subdesign. The data set comprises 517 respondents that were randomly sampled from the Dutch general population.

The same priors that were used in the design optimizations of the TPC-QALY designs were also used for the D-error and sample size evaluations of the designs. However, whereas the design optimizations were based on a full ranking of the choice options, the design evaluations were based on the D-error of the actual presentation format of the matched pairs (ie, A vs B and B vs C, with options A and B randomized). This is computationally more demanding and hence not used for the design optimizations, but it results in more accurate sample size estimates that are not inflated by an auxiliary choice task that is not intended to be seen by respondents. Finally, all sample size estimates were obtained using the calculations as described by De Bekker-Grob et al[21] assuming 80% power and 0.05 significance level.

Impact of Number of Bayesian Draws

Ideally, a small number of draws is specified in the design optimizations because the optimization time scales linearly with the number of draws selected. However, a sufficiently large number of draws needs to be used to ensure that the Bayesian D-error criterion correctly identifies the relative efficiency of small design changes made by the optimization algorithm. Interestingly, the absolute accuracy of the D-error is unimportant; all that is required to obtain the same level of accuracy is that the ranking of the D-error of different designs remains correctly identified.[22] Moreover, given a limited optimization time and a large design space, it can even be optimal to sacrifice some accuracy to be able to perform more optimization iterations in a fixed amount of time.

To evaluate the impact of the number of Bayesian draws on the optimization of TPC-QALY designs, Latin hypercube samples with 50, 100, 200, and 300 draws were optimized using a columnwise-pairwise algorithm, and a fourth sample with 1000 draws was optimized using a genetic algorithm (cf Liefvendahl and Stocki).[23] Then, an exponential TPC-QALY design was optimized using the default TPC-QALY settings as described earlier and using 100 Bayesian draws. During the design optimization, the successive DCE designs and the number of design improvements (ie, the number of choice tasks swapped by the modified Fedorov algorithm) were saved every 10 000 iterations. The D-errors of the saved designs were subsequently evaluated using the different numbers of Bayesian draws, and the rank of the calculated D-errors was calculated to determine whether a different number of draws would have produced divergent rankings of the consecutive designs.

Results

Table 1 presents the results obtained from the D-efficiency comparisons based on the 8 DCE designs. In terms of relative design efficiencies, there is a major difference in statistical performance between the 4 DCE designs that are explicitly optimized for the estimation of time-preference-corrected QALY tariffs and the other 4 conventional DCE designs that are not. As shown, any DCE design optimized for nonlinear time preferences works well irrespective of the choice of discount function. This implies that there is little added value in optimizing a DCE design for multiple discount functions simultaneously, particularly when considering the additional run time that is required for optimizing such designs. In contrast, the conventional DCE designs have considerably lower design efficiencies: the DCE design optimized for a linear multiplicative utility function has an average relative design efficiency of 0.91, the heterogeneous linear additive design with severity stratification 0.80, and the homogeneous linear additive designs without severity stratification 0.69.

Table 1. DCE design efficiency for time-preference-corrected QALY tariff estimations*

Design optimized for:	Absolute and relative Bayesian D-error achieved for:								
	Exponential TPC utility function			Hyperbolic TPC utility function			Power TPC utility function		
	abs.	rel.	rank	abs.	rel.	rank	abs.	rel.	rank
1. Exponential utility function	0.0619	1.00	1	0.0564	1.00	2	0.1844	1.00	3
2. Hyperbolic utility function	0.0621	1.00	2	0.0563	1.00	1	0.1846	1.00	2
3. Power utility function	0.0624	0.99	4	0.0566	0.99	4	0.1846	1.00	1
4. Exponential and hyperbolic and power	0.0622	1.00	3	0.0564	1.00	3	0.1853	1.00	4
5. Linear multiplicative	0.0682	0.90	5	0.0611	0.91	5	0.1997	0.92	5
6. Linear additive	0.0744	0.80	6	0.0674	0.80	6	0.2192	0.81	6
7. Standard linear additive†	0.0807	0.70	8	0.0747	0.67	7	0.2380	0.71	7
8. Standard linear additive 0 priors†	0.0805	0.70	7	0.0751	0.67	8	0.2391	0.70	8

TPC indicates time-preference-corrected; abs, absolute; rel, relative.
*Absolute design efficiencies are based on the Bayesian D-error criterion, and relative design efficiencies are in comparison to the most efficient discrete choice experiment (DCE) designs.
†The standard designs are homogeneous DCE designs without severity stratification; all other designs are heterogeneous DCE designs with severity stratification.

Table 2 presents the asymptotic standard error (SE) and sample size estimates for the discount rate parameters as determined for each of the 8 DCE designs. The 4 DCE designs that were specifically optimized for the estimation of time-preference-corrected QALY tariffs again perform very similarly to one another. They also perform better in terms of SE and sample size estimates than the DCE designs based on a linear utility specification. Similar to the results presented in Table 1, the multiplicative design performs better than the heterogeneous linear additive design with severity stratification and much better than the homogeneous linear additive designs without severity stratification. In fact, the latter designs results in 2.5 to 3.5 times larger SE and up to 11 times larger sample size estimates than those of the nonlinear multiplicative DCE designs.

Table 2 also highlights the implications of the choice of discount function in terms of the statistical identification of the discount rate parameters and recommended minimum sample size. As shown, the asymptotic SE of the hyperbolic discount rate parameter is twice as large, and that of the power function rate parameter 4 times as large as the asymptotic SE of the exponential discount rate parameter. Hence, based on DCE designs that are, in turn, based on the priors as obtained using the data set of Lim et al,[19] the exponential discount function appears to be better identified than the hyperbolic and power discount functions. This translates into larger sample sizes as required for the hyperbolic and power discount functions. Interestingly, the larger SE of the power function is mitigated by the larger power discount rate parameter of 0.33, which is further away from 0 and thus requires a smaller sample to get statistically significant estimates than the exponential and discount rate parameters of 0.11 and 0.12, respectively. This also explains why the power discount function has smaller sample size estimates than the hyperbolic discount function even though its SE of the discount rate parameter is almost twice as large.

Table 2. Asymptotic standard error and sample size estimate for the discount rate parameter*

Design optimized for:	Standard error (SE) and sample size estimate achieved for:								
	Exponential TPC utility function			Hyperbolic TPC utility function			Power TPC utility function		
	SE	sample	rank	SE	sample	rank	SE	sample	rank
1. Exponential utility function	0.43	100	1	0.96	400	1	1.74	175	1
2. Hyperbolic utility function	0.45	108	2	0.99	421	3	1.79	184	4
3. Power utility function	0.46	112	4	1.00	431	4	1.79	184	3
4. Exponential and hyperbolic and power	0.46	110	3	0.99	419	2	1.77	180	2
5. Linear multiplicative	0.58	181	5	1.26	677	5	2.24	289	5
6. Linear additive	0.61	198	6	1.44	885	6	2.66	407	6
7. Standard linear additive†	1.11	650	7	2.67	3055	7	5.00	1440	7
8. Standard linear additive 0 priors†	1.42	1068	8	3.21	4435	8	6.04	2095	8

TPC indicates time-preference-corrected.
*Estimated sample size to determine significant deviation from linear time preferences with $\alpha = 0.05$ and 80% power.
†The standard designs are homogeneous discrete choice experiment (DCE) designs without severity stratification; all other designs are heterogeneous DCE designs with severity stratification.

Figure 1. Bayesian D-errors of designs optimized for an exponential TPC utility function. * All Bayesian draws were created using pre-optimized Latin hypercube samples (LHS) and based on the priors for the exponential TPC utility function. The total optimization time for 200 000 iterations with 100 Bayesian draws was 15 minutes (single threaded) on an Intel core i7-8086k.

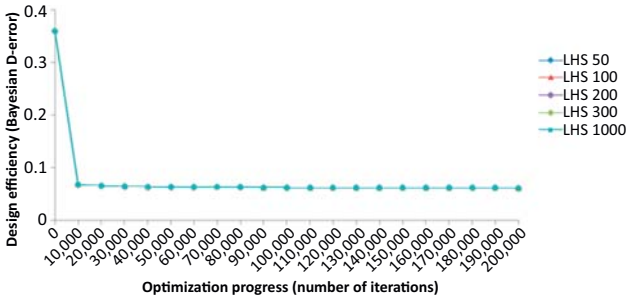

Figure 1 provides an overview of the design optimization progress for the TPC-QALY designs. As shown, the optimization algorithm achieves the vast majority of the design improvements within the first 10 000 optimization iterations: 98% of the total decrease in the D-error was achieved in the first 10,000 optimization iterations, 99% within the first 30 000 iterations, and less than 1% in the remaining 170 000 iterations. More important, irrespective of the number of Bayesian draws, all of the design iterations are ranked identically (note that more detailed output is presented in Appendix Table 2 in Appendix A in the Supplementary Materials). Hence a larger number of Bayesian draws did not increase the optimization accuracy.

Discussion

The estimation of time-preference-corrected QALY tariffs requires a multiplicative utility function and involves a nonlinear discount function, which are not supported by existing DCE optimization packages. As shown in this article, standard DCE designs that can be optimized using currently available software packages are inefficient, which implies that substantially larger sample sizes are required to obtain a similar level of reliability. For this reason, we strongly recommend the use of appropriately optimized DCE designs for the estimation of time-preference-corrected QALY tariffs.

A potential limitation of the results presented is that they are based on a specific set of priors, which were based on a single data set that, in turn, was obtained using a DCE design that was not specifically optimized to accommodate any of the included discount functions. Additionally, the data set that was used to obtain the priors comprised more than 500 respondents and consequently resulted in relatively precise conditional logit estimates. For this reason, we present in Appendix B (in the Supplementary Materials found at https://doi.org/10.1016/j.jval.2019.05.014) the results of a sensitivity analysis in which the size of the standard errors of the priors was substantially increased. The size of the standard deviations was maximized under the constraint that 99.5% of the Bayesian draws would retain the same sign as the prior mean and thus retain the "correct" sign in the design optimizations. These standard deviations also correspond to the maximum amount of preference heterogeneity that is allowed by the TPC-QALY program. Accordingly, it is comforting to establish that the results presented in Appendix B closely correspond to the results presented in the main text.

In addition to introducing the Bayesian D-error optimization criteria and evaluating the relative performance of designs created using these criteria, this article is accompanied by an easy-to-use software implementation that can generate DCE-duration designs for the most commonly used exponential, hyperbolic, and power discounting functions. Our software package, called TPC-QALY, optimizes for the correct D-efficiency criterion, includes attribute-level overlap on both the duration and health attributes, supports heterogeneous DCE designs, and automatically applies health-state severity stratification. Accordingly, the TPC-QALY software allows applied researchers to easily generate theoretically sound and efficient DCE designs that accommodate the estimation of time-preference-corrected QALY tariffs.

At the same time, there is an important caveat. The design optimization crucially relies on informative priors to implement the severity-stratified candidate sets, to determine the selected health states, and to optimize the included duration values. For this reason, the TPC-QALY software package is not intended to be used with uninformative priors. Furthermore, because the severity stratification and the selection of duration values in the DCE design will be sensitive to the specified discount values, we recommend that 1 or more pilot samples are used to update the priors during the data collection. When doing so, the presented results suggest that the exponential discount function produces pilot results with the smallest standard errors. From this perspective, it is comforting that the presented results also confirm that the optimization of TPC-QALY designs with one type of discount function does not preclude the estimation of time-preference-corrected QALY tariffs with 1 of the alternative discount functions.

Acknowledgments

We gratefully acknowledge financial support from the EuroQol Research Foundation.

The views and opinions expressed in this article are those of the authors and do not necessarily reflect those of the EuroQol Group.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2019.05.014.

REFERENCES

1. Jonker MF, Donkers B, de Bekker-Grob EW, Stolk EA. Advocating a paradigm shift in health-state valuations: the estimation of time-preference corrected QALY tariffs. *Value Health*. 2018;21(8):993–1001.
2. Soman D, Ainslie G, Frederick S, et al. The psychology of intertemporal discounting: Why are distant events valued differently from proximal ones? *Market Lett*. 2005;16(3-4):347–360.
3. Attema AE, Brouwer WB. On the (not so) constant proportional trade-off in TTO. *Qual Life Res*. 2010;19(4):489–497.
4. Jakubczyk M, Craig BM, Barra M, et al. Choice defines value: a predictive modeling competition in health preference research. *Value Health*. 2018;21(2):229–238.
5. Craig BM, Rand K, Bailey H, Stalmeier PF. Quality-adjusted life-years without constant proportionality. *Value Health*. 2018;21(9):1124–1131.
6. Jonker MF, Attema AE, Donkers B, Stolk EA, Versteegh MM. Are health state valuations from the general public biased? A test of health state reference dependency using self-assessed health and an efficient discrete choice experiment. *Health Econ*. 2017;26(12):1534–1547.
7. Mulhern B, Bansback N, Hole AR, Tsuchiya A. Using discrete choice experiments with duration to model EQ-5D-5L health state preferences: testing experimental design strategies. *Med Decis Making*. 2017;37(3):285–297.

8. Versteegh MM, Vermeulen KM, Evers SM, de Wit GA, Prenger R, Stolk EA. Dutch tariff for the five-level version of EQ-5D. *Value Health*. 2016;19(4):343–352.
9. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: An EQ-5D-5L value set for England. *Health Econ*. 2018;27(1):7–22.
10. Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Qual Life Res*. 2011;20(10):1727–1736.
11. Samuelson PA. A note on measurement of utility. *Rev Econ Stud*. 1937;4(2):155–161.
12. Mazur JE. An adjusting procedure for studying delayed reinforcement. In: Commons ML, Mazur JE, Nevin JA, Rachlin H, eds. *Quantitative Analyses of Behavior*. Vol 5. Hillsdale, NJ: Erlbaum; 1987:55–73.
13. Stevens SS. On the psychophysical law. *Psychol Rev*. 1957;64(3):153–181.
14. Scarpa R, Rose JM. Design efficiency for non-market valuation with choice modelling: how to measure it, what to report and why. *Aust J Agric Resource Econ*. 2008;52(3):253–282.
15. Kanninen BJ. Optimal design for multinomial choice experiments. *J Market Re*. 2002;39(2):214–227.
16. Rose JM, Bliemer MC. Constructing efficient stated choice experimental designs. *Transport Rev*. 2009;29(5):587–617.
17. Jonker MF, Donkers B, de Bekker-Grob EW, Stolk EA. Attribute level overlap (and color coding) can reduce task complexity, improve choice consistency, and decrease the dropout rate in discrete choice experiments. *Health Econ*. 2019;28(3):350–363.
18. Jonker MF, Donkers B, de Bekker-Grob EW, Stolk EA. The effect of level overlap and color coding on attribute non-attendance in discrete choice experiments. *Value Health*. 2018;21(7):767–771.
19. Lim S, Jonker MF, Oppe M, Donkers B, Stolk EA. Severity-stratified discrete choice experiment designs for health state evaluations. *Pharmacoeconomics*. 2018;36(11):1377–1398.
20. Sándor Z, Wedel M. Heterogeneous conjoint choice designs. *J Market Res*. 2005;42(2):210–218.
21. de Bekker-Grob EW, Donkers B, Jonker MF, Stolk EA. Sample size requirements for discrete-choice experiments in healthcare: a practical guide. *Patient*. 2015;8(5):373–384.
22. Kessels R, Jones B, Goos P, Vandebroek M. An efficient algorithm for constructing Bayesian optimal choice designs. *J Business Econ Stat*. 2009;27(2):279–291.
23. Liefvendahl M, Stocki R. A study on algorithms for optimization of Latin hypercubes. *J Sta Planning Inference*. 2006;136(9):3231–3247.