



Preference Paths and Their Kaizen Tasks for Small Samples

Benjamin Matthew Craig¹ · Kim Rand^{2,4} · John D. Hartman³

Accepted: 15 July 2021 / Published online: 30 July 2021

© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2021

Abstract

Background Stated preference research currently lacks a form of evidence that is well suited for small samples. A preference path is a sequence of two or more choices showing the evolution of an object following an adaptive process.

Objectives The aims were to introduce preference paths and their kaizen tasks and to demonstrate how to analyze their evidence using a small sample.

Methods Twenty respondents were assigned the same 16 profiles generated from an orthogonal array based on the five attributes of the EQ-5D-5L descriptive system. Each kaizen task began with an opt-out paired comparison (i.e., choosing between the initial 10-year profile and the opt-out “dying immediately”), followed by choosing three changes, and ended with a second paired comparison (final profile versus opt-out) if the respondent chose opt-out initially. By maximum likelihood with respondent clusters, we estimated the 20 main effects using conditional logit and Zermelo–Bradley–Terry (ZBT) specifications.

Results Apart from demonstrating heterogeneity and profile effects, all main effect estimates were non-negative, and most were significant (15 for logit and all 20 for ZBT; p value < 0.05). Under the logit and ZBT specifications, the value of the worst EQ-5D-5L profile (55555) is – 0.920 quality-adjusted life years (QALYs) or – 1.478 QALYs, respectively. Furthermore, the findings illustrate a log-linear relationship between the logit and ZBT main effects.

Conclusion This paper demonstrates the feasibility of a stated-preference study that estimates 20 main effects using path evidence from 20 respondents (16 kaizen tasks, 15-min interviews). This approach shows promise for future application in stated-preference research, particularly in small samples.

Key Points for Decision Makers

Eliciting a sequence of preferences along a pathway offers a novel approach for stated-preference researchers, particularly when faced with small samples.

This study demonstrates how to implement this adaptive task and estimate 20 main effects using preference evidence collected from 20 respondents during 15-min interview surveys.

Its results show that the estimates produced using the Zermelo–Bradley–Terry (ZBT) and logit models have a log-linear relationship. Unlike the logit, the ZBT estimates do not require scaling parameters or additional constraints, which is particularly advantageous in health valuation.

✉ Benjamin Matthew Craig
bcraig@usf.edu

¹ Department of Economics, University of South Florida, Tampa, Florida, USA

² Health Services Research Center, Akershus University Hospital, Lørenskog, Norway

³ Department of Health Science and Administration, University of West Florida, Pensacola, Florida, USA

⁴ Maths in Health B.V., Rotterdam, The Netherlands

1 Introduction

Discrete choice experiments (DCEs) are often conducted to elicit stated preferences from respondents. Given a choice set and a hypothetical scenario, each decision maker selects the alternative that maximizes their utility [1]. Such discrete behaviors imply inequalities (e.g., $A > B$) that resolve ambiguities between objects [2]. The number and diversity of alternatives in a choice set determines the breadth of preference evidence. However, too much breadth may reduce response quality because the tasks involve so many alternatives or trade-offs that they confuse respondents or exceed the capacity of respondents' working memories [3, 4]. Apart from mimicking the decision context, researchers must design each preference elicitation task carefully, keeping breadth and response quality in mind.

At one extreme, some researchers maximize response quality in their DCE by minimizing the set size or the number of differential attributes. They may compensate for the paucity of breadth by increasing the number of respondents. Such large samples may be affordable with online surveys of the general population. For example, the authors previously conducted a national valuation study with 3160 pairs and 8222 respondents [5]. However, this brute force approach is not feasible when the sample size is limited to 100 respondents or fewer (e.g., patient sample) or the descriptive system naturally entails five or more attributes (e.g., treatment decisions).

When large samples with few attributes are not feasible, some valuation studies elicit indifference statements instead of inequalities. If their assumptions hold, indifference statements imply that the difference in utility between two objects is small (i.e., $|A - B| < \epsilon$) [2]. In other words, when respondents are asked, "Which do you prefer?" between two objects, they respond, "I don't know," suggesting that the two objects are of nearly equal utility. Valuation studies commonly use adaptive paired comparisons to produce indifference statements [2]. Such tasks adapt along an attribute, known as a numeraire.

For example, the time trade-off (TTO) task adjusts the time attribute of one object until its value is within range of its counterpart [6]. The numeraire may be time (TTO), money (willingness to pay [7]), risk (standard gamble [6]), or persons (person trade-off [PTO] [8]); yet, each task implies that every respondent comprehends the concepts underlying the numeraire, wants more or less of the numeraire, is willing to trade the numeraire for something else, and shares a common definition of "near equivalence." Many studies have rejected these assumptions [5, 9, 10]; however, the limitations of indifference statements and their threshold tasks have been tolerated due to the absence of an alternative approach for use in small samples.

1.1 Preference Paths

In this paper, we propose an alternative form of preference evidence that is particularly well-suited for small samples. Its terminology may be unfamiliar to some readers (see the "Glossary"; defined terms are italicized at their first appearance in the text). For example, a preference path is a sequence of two or more choices showing the evolution of an object following an adaptive process. Imagine a descriptive system with five five-level attributes in which one is the best attribute level and five is the worst attribute level (e.g., EQ-5D-5L). In this system of five attributes, the objects' profiles range from 55555 (with each attribute at their worst level) to 11111 (with each attribute at their best level). To elicit a preference path (Fig. 1), a respondent might start with a single profile 33333 and be presented with a choice set to include gains in each attribute by one level. If the respondent chooses to improve the fifth attribute, the profile changes to 33332. Next, the respondent is presented with a choice set with gains in each of the four remaining attributes by one level, and so on. With this particular format, the path described by the respondent's choices reveals the evolution of an object's profile (shown in red in Fig. 1) from 33333 to 22222.

Although Fig. 1 shows a path with one-level gains, alternative tasks may ask respondents to choose between losses in levels, multi-level gains/losses, or gains/losses in multiple attributes. While we developed the concept of preference paths for this project, three econometric adjustments emerged that greatly facilitated its execution, creating the *kaizen task*. Apart from introducing the *kaizen task* and the three adjustments, we demonstrate methods for analyzing preference paths using empirical data from a small-sample EQ-5D-5L valuation study. Its interview script and de-identified data are also included as electronic supplemental material. With further development, preference paths and their *kaizen tasks* may become a commonplace in stated preference research, particularly for small samples.

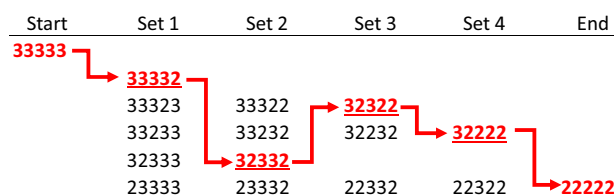


Fig. 1. Example of a *kaizen task* and a preference path

2 Methods

2.1 Kaizen Tasks

Kaizen is a Japanese term describing continuous improvement (derived from two Kanji, the first 'Kai' 改, meaning 'change,' and the second 'zen' 善, meaning 'good'), which, in this case, is the discrete evolution of an object over a sequence of choices. As introduced by the co-authors in this paper, each choice along a preference path reflects the preferred change to the object's profile given the choice set. This novel task relies on a single profile of attribute levels taken from a well-defined descriptive system, making it easy to understand and quick to complete, particularly online. Based on qualitative feedback (see "Acknowledgments"), respondents enjoy the evolving format because it gives them control to modify the profile as they see fit. This interactive experience is different from non-adaptive tasks, where each respondent makes a single choice and does not see its implications.

The process of completing a kaizen task is familiar akin to decorating a model (e.g., Mr. Potato Head [an American toy consisting of a plastic model of a potato "head" to which a variety of plastic parts can attach—typically ears, eyes, shoes, hat, nose, pants, and mouth]) or developing a character for a game, where players trade attributes in a profile (e.g., Diablo [an action role-playing video game originally released in 1997], Civilization [a turn-based strategy video game originally released in 1991]). Given an initial profile, a person makes sequential choices to improve an object, revealing a preference path.

Each kaizen task embodies the Markov property in that the likelihood of each choice is between zero and one and depends only on the choice set, not on any previous choice. Therefore, each choice along the preference path is stochastic and implies one or more inequalities. Two conditions are necessary for this property to hold.

First, each choice set must include two or more different alternatives. This implies that the attributes in the descriptive system may not share seemingly identical levels. For example, a pain attribute may include "no pain" as its best level. Likewise, a discomfort attribute may include "no discomfort" as its best level. The similarity of these levels (no pain vs. no discomfort) may complicate the interpretation of the choice for some respondents. As an extreme example, if the two levels are clearly identical, the choice set inherently includes only one alternative, violating the Markov property (i.e., likelihood is one).

Second, each choice must not allow the modification of any previously modified attribute. Allowing an attribute to be modified more than once may confound the preference evidence because the choice of the first modification

may be motivated to permit the subsequent modification. The respondent does not consider the past or future once the choice set is present. Combined, these two conditions imply that the maximum number of choices per task is the number of attributes minus one.

The initial version of the kaizen task shown in Fig. 1 involves four choices and produces ten inequalities of preference evidence. However, an alternative task might include a "no change" alternative, reducing the number of inequalities and increasing the number of possible paths. In this kaizen task, each subsequent choice has a smaller set size. The initial choice may seem difficult due to the large set size, but the decision is easier because preference intensity (i.e., magnitude of attribute importance) is also large. Later choices may seem easier due to smaller set sizes, but these decisions may in some cases increase in difficulty as preference intensity declines with each choice. Recognizing this decline, the number of choices in a kaizen task may be limited using stopping criteria (e.g., only three changes). Alternatively, the adaptive process could put attributes "in play" after each choice to stabilize the set size. For example, after choosing between three potential changes, the respondent may choose between the two remaining and a new one. This is akin to card games (e.g., gin rummy or poker), where players may discard and draw equal numbers of cards, keeping the number of cards in their hand fixed and causing their decision making to be more manageable.

2.2 Three Adjustments that Enhance Kaizen Tasks

In Fig. 1, each choice entails a one-level change of an attribute. While this initial version was attractive because it is easy to understand and perform, its reliance on trading one-level changes creates a perfect collinearity between the main effects, making it impossible to identify whether a choice is made in favor of one attribute level or against its counterparts. Within the econometric analysis, a researcher may constrain an attribute level to zero so that the remaining coefficients represent a difference in the relative importance of attribute levels. However, this approach impedes probability prediction and scaling.

As a practical solution, we first adjusted the kaizen task to start and end with a non-adaptive task, namely an *opt-out* paired comparison (initial profile vs. opt-out). Instead of only trading off single-level gains, the initial and final paired comparisons draw attention to the full profiles holistically and ask the respondent whether the opt-out is preferred before and after their changes. The resulting combination of evidence promotes the coefficient identification, probability prediction, and scaling as shown in the EQ-5D-5L valuation study.

The second adjustment relates to the use of nominal attributes. For any specific respondent, the levels of an

attribute are ordered, but this order may vary between respondents. In some kaizen tasks, one respondent may perceive a choice between two potential gains, and another, faced with the same set, sees only one. To avoid violating the first condition of the Markov property for some respondents, the maximum number of potential changes per kaizen task is the number of ordinal attributes minus one.

Alternatively, a researcher may be tempted to include a “no change” alternative and a stopping criterion known as a threshold. If respondents do not perceive any merit in the potential changes, they may choose “no change” or say “I don’t know,” causing the task to terminate. However, offering such a stopping criterion leads to the endogenous censoring of the preference evidence (i.e., favoring non-response over the stated preference). For example, the respondents have a direct incentive in a TTO to report indifference prematurely, because it ends the task faster.

The final adjustment relates to the use of *hold-outs*. A hold-out is an attribute that is common to all alternatives within a choice set. For example, a kaizen task that trades off potential changes may start with an ordinal attribute at its best level. Since the attribute cannot be improved, it is common to all alternatives in every set. Therefore, a kaizen task might impose a correlation between levels and hold-outs inadvertently. To dissolve this correlation, a researcher may impose hold-outs at inferior levels, expressing the level of an attribute and disallowing its improvement. Hold-outs are also useful when interactions are hypothesized. Regardless, each hold-out that is imposed decreases the maximum number of potential changes per kaizen task.

In combination, the initial kaizen task (Fig. 1) was adjusted in three ways: it starts and ends with an opt-out paired comparison; the number of potential changes is limited to the number of ordinal attributes minus one; and initial profiles with attributes at their best levels were not permitted (i.e., no hold-outs). These three adjustments are demonstrated in the EQ-5D-5L valuation study.

2.3 The EQ-5D-5L Valuation Study

This study was designed to demonstrate a kaizen task, accounting from the EQ-5D-5L descriptive framework and the three econometric adjustments. Its framework includes a hypothetical scenario and a descriptive system of five attributes (see the “Interview Materials” [pdf file] in the electronic supplementary materials): Mobility (MO), Self-Care (SC), Usual Activities (UA), Pain/Discomfort (PD), and Anxiety/Depression (AD). The hypothetical scenario is, “Starting today, you could have the following health problems for the next 10 years, then die.” The paired comparisons at the start and end of each kaizen task included “dying immediately” is an opt-out.

Each attribute was characterized as a health problem being at 1 of 5 levels: none (level 1), slight, moderate, severe, unable/extreme (level 5). Generally, their levels are considered ordinal; however, previous studies have shown that some respondents may reverse their order due to variation in the interpretation of specific words describing the magnitude of problems, such as “severe” and “extreme,” particularly for PD and AD [11, 12]. Therefore, the maximum number of potential changes per kaizen task is three in this descriptive system.

As a highly convenient sample, the two co-authors (KR and JH) and 18 colleagues who have previously collaborated with the authors were interviewed by BC (see the “Acknowledgements”). Each respondent was assigned the same 16 profiles in the same order. The profiles were generated from an orthogonal array: 22222, 23333, 24444, 25555, 32345, 33254, 34523, 34523, 35432, 42453, 43542, 44235, 45324, 52534, 53425, 54352, 55243. They did not include any attributes at level one or other hold-outs.

Each interview (in-person or video conference) began by reading the script aloud to the respondent (see the “Interview Materials” [pdf] in the electronic supplementary materials). Each of the 16 tasks began with an opt-out paired comparison (i.e., a choice between the initial profile and opt-out), followed by choosing three potential changes. Like Fig. 1, the first choice set had five alternatives, the second had four, and the third had three, creating 60 possible paths ($5 \times 4 \times 3$) from each initial profile. At the end of the task, the respondents who chose to opt out initially were asked to complete a second opt-out paired comparison (i.e., final profile vs. opt-out). This final response indicates whether the three changes altered the object’s profile sufficiently to persuade the respondent to opt in and not “die immediately.” BC recorded the response vector on a spreadsheet (see the “Data” [csv file] in the electronic supplementary materials). Generally, each interview (16 tasks) took less than 15 min; however, respondents typically provided qualitative feedback and discussed the task design with BC afterwards.

2.4 Econometric Analysis

The analysis plan entailed a descriptive analysis of response quality to characterize internal validity and a primary analysis of the main effects using two alternative specifications for choice modeling. Each analysis was conducted using Stata 14 software [13].

To model the choice probabilities, we estimated a conditional logit model, expressed as $\Pr(y_{ij} = 1) = \frac{\exp(X_{ij}\beta)}{\sum_{k=1}^J \exp(X_{ik}\beta)}$, and a Zermelo–Bradley–Terry (ZBT) model, expressed as $\frac{X_{ij}\beta}{\sum_{k=1}^J X_{ik}\beta}$ [5, 10, 14–17]. In each model, the regression $X_{ij}\beta$ describes the causal relationship between the differential attribute levels and the likelihood of a choice, given its

alternatives. It includes a constant and 20 incremental dummy-coded variables, four for each of the five attributes. To aid interpretation, the constant is constrained to one, which inherently divides each coefficient by the difference in value between “10 years with no health problems then die” and “dying immediately.” Within the health valuation literature, this difference is known as 10 quality-adjusted life years (QALYs). Therefore, each coefficient is expressed as a loss on a QALY scale and was estimated by maximum likelihood with respondent-specific clusters.

To adjust for the proportional difference between the In-odds and QALY scales, the logit model includes a scaling parameter α such that $\beta_k = \alpha \times \beta_k^*$. Under the ZBT model, the scale parameter α is not necessary because it is cancelled in the ratio [17]. For both models, the parameters were estimated by maximum likelihood with respondent clusters, which accommodates multiple responses per respondent. Like all other health valuation studies, we hypothesized that each parameter ($\beta_1^*, \dots, \beta_{20}^*$ and α) is positive. The significance of each parameter was assessed using a percentile bootstrap with replacement, respondent-specific clusters, and 1000 iterations.

3 Results

3.1 Response Behaviors

All 20 respondents were assigned the same 16 initial profiles in the same order. Two respondents (10%) never chose opt-out initially, and nine (45%) never chose opt-out at the end of their tasks, suggesting heterogeneity in the value of “dying immediately.” Out of the 16 tasks, two initial profiles and seven final profiles were unanimously better than “dying immediately,” suggesting profile effects.

Among the 60 unique orders in which attributes could be selected, 53 (88%) were reported. In particular, the first attribute chosen varied: 35% PD, 30% AD, 15% SC, 14% UA, and 6% MO. In each of the 16 tasks, no attribute was chosen first by all 20 respondents, suggesting response heterogeneity. No respondent began all tasks by improving the same attribute, suggesting profile effects. Each respondent tailored their preference paths based on the initial profile, reporting between 8 and 15 unique attribute orders.

3.2 Two Modifications in the Logit Estimation

As proposed originally, the primary analysis included estimations of the logit and ZBT models (Table 1). However, the logit estimation failed to converge, calling for two deviations. First, separate scale parameters α were estimated for the opt-out paired comparisons (3.567; 95% confidence interval [CI] 2.589–5.276) and the preference path (34.945;

Table 1 Main effects for the conditional logit and Zermelo–Bradley–Terry (ZBT) specifications

	Conditional logit ^a		ZBT	
	QALY loss (95% CI)	p value	QALY loss (95% CI)	p value
MO1	0.001 (– 0.024, 0.018)	0.990	0.001 (0.000, 0.002)	< 0.001
MO2	0.068 (0.051, 0.083)	< 0.001	0.009 (0.002, 0.018)	< 0.001
MO3	0.116 (0.094, 0.142)	< 0.001	0.052 (0.019, 0.100)	< 0.001
MO4	0.147 (0.123, 0.177)	< 0.001	0.160 (0.076, 0.284)	< 0.001
SC1	0.000	NA	0.001 (0.000, 0.002)	< 0.001
SC2	0.064 (0.047, 0.085)	< 0.001	0.008 (0.002, 0.019)	< 0.001
SC3	0.112 (0.092, 0.139)	< 0.001	0.045 (0.019, 0.078)	< 0.001
SC4	0.135 (0.110, 0.172)	< 0.001	0.102 (0.043, 0.212)	< 0.001
UA1	0.010 (– 0.008, 0.025)	0.300	0.001 (0.000, 0.002)	< 0.001
UA2	0.083 (0.064, 0.102)	< 0.001	0.016 (0.005, 0.030)	< 0.001
UA3	0.130 (0.107, 0.159)	< 0.001	0.086 (0.039, 0.147)	< 0.001
UA4	0.157 (0.128, 0.196)	< 0.001	0.228 (0.107, 0.406)	< 0.001
PD1	0.014 (– 0.012, 0.033)	0.228	0.001 (0.000, 0.003)	< 0.001
PD2	0.098 (0.071, 0.123)	< 0.001	0.026 (0.009, 0.048)	< 0.001
PD3	0.164 (0.128, 0.213)	< 0.001	0.304 (0.110, 0.726)	< 0.001
PD4	0.192 (0.151, 0.255)	< 0.001	0.777 (0.322, 2.216)	< 0.001
AD1	0.014 (– 0.019, 0.038)	0.328	0.001 (0.000, 0.003)	< 0.001
AD2	0.091 (0.061, 0.116)	< 0.001	0.021 (0.007, 0.038)	< 0.001
AD3	0.155 (0.122, 0.195)	< 0.001	0.214 (0.080, 0.452)	< 0.001
AD4	0.173 (0.139, 0.219)	< 0.001	0.425 (0.191, 1.048)	< 0.001

For example, the worst EQ-5D-5L profile (55555) is – 0.920 QALYs or – 1.478 QALYs, respectively. Each value is one minus the sum of the 20 estimates of QALY losses

AD Anxiety/Depression, CI confidence interval, MO Mobility, PD Pain/Discomfort, QALY quality-adjusted life year, SC Self-Care, UA Usual Activities

^aThe logit scale parameter is significantly greater in the paired comparisons (3.567; 95% CI 2.589–5.276) than the preference path (34.945; 95% CI 26.242–49.986)

95% CI 26.242–49.986). Choices between potential changes in a profile are at a greater magnification than the choices between the full profile and opt-out; therefore, the logit model required two scaling parameters. In this case, the scale parameter of the path is a tenth of the paired comparisons, and this difference further motivates the use of the kaizen task.

Second, the logit estimation failed to identify the level-1 coefficients, which, in this case, represented the differences between no problems and slight problems for each attribute. As described in the methods, the opt-out paired comparisons were included in the task to facilitate coefficient identification given the perfect collinearity between potential changes in the preference paths. Based

on the results, the paired comparison evidence was sufficient to identify the severe and extreme coefficients, but not the coefficients for smaller changes, likely because of the minimal effects of small changes on opting out. By constraining the smallest coefficient to zero (SC1) and allowing task-specific scale coefficients, the logit estimation converged. Alternatively, the ZBT model estimation converged without either modification.

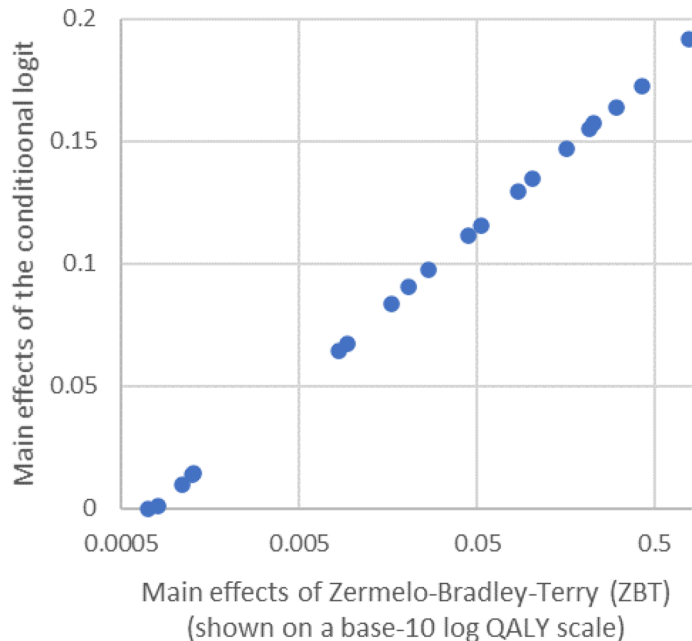
3.3 EQ-5D-5L Valuation

The logit estimation shows that 15 of the 20 potential changes caused significant losses in QALYs (p value < 0.05; Table 1). The ZBT shows that all 20 potential changes were significant. Under the logit and ZBT specifications, the value of the worst EQ-5D-5L profile (55555) is one minus the sum of the 20 estimates of QALY losses, which is -0.920 QALYs or -1.478 QALY, respectively. The ZBT estimates have a log-linear relationship with their logit counterparts (Fig. 2): for all of the potential changes k , the difference between

$0.1970 + 0.0275 \ln(\hat{\beta}_{k,ZBT}^*)$ and $\hat{\beta}_{k,logit}^*$ is less than 0.0025 QALYs.

4 Discussion

This paper introduces kaizen tasks and preference paths for use in stated preference research as well as corresponding terminology (see the ‘‘Glossary’’) and methodological considerations. Using a convenience sample taken from 20 colleagues, the descriptive evidence shows how the task captures preference heterogeneity and profile effects. In the primary analysis of this small sample, we estimated the 20 main effects of the logit and ZBT models and the need for two modifications for logit convergence. Originally, the comparison of alternative specifications was motivated by prior evidence that sigmoidal models, such as logits and probits, have poor predictive validity for health valuation studies due to the heterogeneity in alternatives (i.e., chocolates vs. cars) [5, 10, 14]. These results add to these findings by illustrating a log-linear relationship between their main effects (Fig. 2). In summary, this paper demonstrates the feasibility of a stated-preference study that estimates 20



* Each of the 20 main effects are shown on a QALY scale; however, the x-axis was logged to better illustrate the log-linear relationship between the logit and ZBT estimates.

Fig. 2 Main effects of the conditional logit and Zermelo-Bradley-Terry (ZBT) specifications. Each of the 20 main effects are shown on a quality-adjusted life year (QALY) scale; however, the x-axis was

logged to better illustrate the log-linear relationship between the logit and ZBT estimates

main effects using 20 respondents (16 kaizen tasks, 15-min interviews). This small health preference study will serve as the basis for future work on kaizen tasks and preference paths [18, 19].

Generally, the kaizen task has three potential advantages. First, this adaptive task collects more precise preference evidence than fixed choice tasks (i.e., a tenth of the logit scale parameter), which implies that trading behaviors between potential changes in an object’s profile may better identify causal relationships in small samples and heterogeneous effects within and between groups than choosing between a set of full-profile alternatives.

Second, a kaizen task is not a threshold task and does not rely on a single numeraire, such as time or out-of-pocket costs. For example, a TTO task asks respondents to trade years of life to achieve an indifference threshold [2]. In fact, the concept of a preference path was first hypothesized to account for the respondents’ paths in a TTO [20]. Nevertheless, some TTO respondents have refused to trade life years (i.e., non-trader) or become confused along the way [9]. Protocols typically relied on interviewers to train and guide respondents, but this has led to interviewer effects and other biases [9, 21].

Without the distraction of a numeraire, kaizen tasks have an intuitive appeal for many applications, including health, transportation, and environmental studies. Using a single

profile as the starting point, respondents may find it easier to visualize potential changes than when making a discrete choice between similar alternatives. Instead of relying on a single choice, the preference path promotes greater personalization over a series of choices and likely offers greater resemblance to the real-world decisions that individuals must make (i.e., given your budget, which change would you choose first?). This evolution toward an ideal is particularly advantageous for sensitive topics where respondents do not want to choose between two bad alternatives (e.g., pollution vs. longer commutes), but appreciate the opportunity to pursue a laudable goal [22].

The third advantage is the format’s simplicity, presenting the same amount of information as a paired comparison and evolving the profile between choices. Simple tasks like the one presented here have gained in popularity because they can be conducted online, reaching a broader geographic population and producing quick results without interviewers and with a fraction of the resources. Furthermore, past studies have shown that choice tasks like this one can be conducted on a variety of devices, including smartphones and tablets [23, 24].

Overall, kaizen tasks reside within the family of single-profile tasks, such as case-2 best–worst scaling (BWS). Each single-profile task starts with a description of a single object (instead of showing multiple objects) and elicits responses



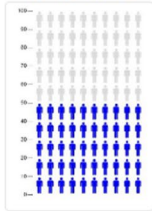
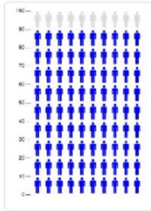
Warmup task:	Vaccination	Potential changes
Duration of immunity	Three months	Six months
Proof of vaccination	No vaccination card	Vaccination card
Vaccination setting	Community setting	Medical setting
Risk of severe side effects	Moderate risk: 1 side effect in 1,000 	Lowest risk: 1 side effect in 1,000,000 
Vaccine effectiveness	50% Effective 	90% Effective 
2a. Which potential change do you choose first?	Vaccination	Potential changes

Fig. 3 Two-column format of a kaizen task [25]

regarding its profile, such as an object's attributes or potential changes to its profile. For example, the kaizen task in Fig. 3 is from a recent online study of coronavirus disease 2019 (COVID-19) vaccination preferences that used a two-column format [25]. Looking at the profile in its first column, a case-2 BWS task might ask, "Which is the worst attribute?" [26]. However, this attribute may or may not be the one that you want to change first in a kaizen task. Identifying the worst attribute does not directly imply preferences regarding potential changes. In a kaizen task, the attention is drawn to the *potential changes* in an object's profile, not the attribute levels alone. Also, each response in a kaizen task represents a choice from a set of nearly identical objects. Unlike case-2 BWS evidence, preference paths may be analyzed using methods common to discrete choice modeling, as shown in this paper.

Among the single-profile tasks, a kaizen task may be considered an adaptive form of a pivoted task [27, 28]. In stated-preference research, a pivot is an object used across choice tasks, sometimes as a base case modified within a choice set. For example, a task may introduce a full profile with 30 attributes and ask the respondent to choose between three partial profiles, breaking the full profile into manageable pieces. A pivoted task omits descriptions of the attributes shared by the alternatives and is typically not adaptive. A kaizen task displays all attribute levels of the single profile and allows the profile to evolve with each choice (i.e., continuous improvement). Pivoted and kaizen tasks are particularly useful when a descriptive system includes more attributes than a respondent's memory can retain. Future research may compare these tasks in various contexts.

The display format of a kaizen task may be extended to include the entire descriptive system, revealing all possible objects. For example, the attributes of the EQ-5D-5L descriptive system share a similar structure of five levels, and each object may be presented using a grid (Fig. 4). For a kaizen task, the initial profile is shown using full circles, and the potential changes are presented as open ones. In this

grid format, the respondents prioritize the potential changes while seeing the full descriptive system as a guide. With some adjustments, the format could be expanded to allow for items with diverging level structures or that are adapted to various screen sizes. Unlike the two-column format, the grid is constant between tasks; only the locations of the full and open circles change. A future study may test for differences between the two-column and grid tasks within various descriptive frameworks.

The primary limitation of preference path evidence is the perfect collinearity between trade-offs. As shown by these results, this limitation can be overcome by incorporating paired comparisons at the start and end of the kaizen task and either estimating the ZBT model or adjusting for differential scales in the logit model. A second potential limitation is that the adaptation of an object's profile implies endogeneity in the set assignment. Respondents may be randomized to initial profiles drawn from an orthogonal array, but the subsequent profiles depend on prior responses, which may lead to biased estimates. For example, the worst attribute levels were included less frequently because they were the first to be eliminated along the preference paths. Future research may assess the extent of this bias and how to improve the experimental design.

Lastly, the authors' colleagues generously provided a convenient sample of responses to demonstrate preference paths and their kaizen task in health valuation. These 20 esteemed colleagues include two of the authors. Each agreed to be named in the acknowledgments and we thank them again for their time and feedback. The worked example is a methodological demonstration and should not be confused with a typical survey, which would require ethical approval. Studies examining more practical examples, such as US COVID-19 vaccination uptake (Fig. 3), have recently been fielded [25] and will build on these findings.

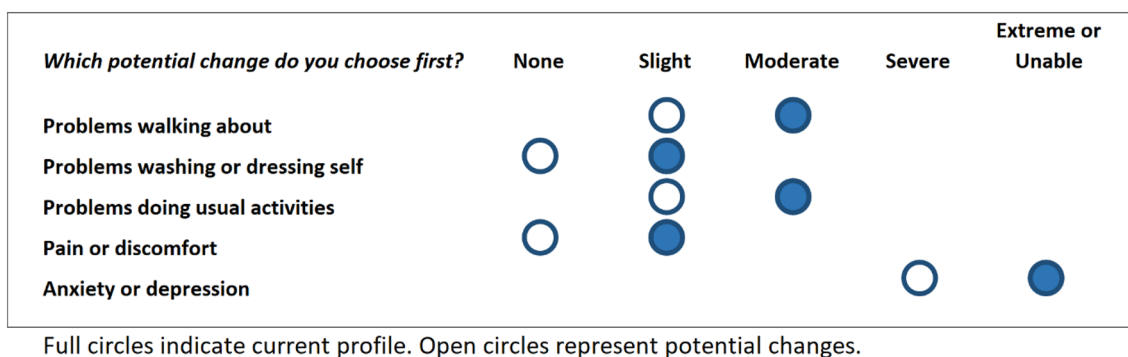


Fig. 4 Grid format of a kaizen task

5 Summary

Preference paths emphasize trade-offs between potential changes to a single profile (Fig. 1). Its kaizen tasks empower respondents to define continuous improvement and do not rely on a numeraire, threshold, or partial profile (Figs. 3, 4). By presenting only a single profile and its potential changes, the task format reduces cognitive burden and mitigates potential sources of framing effects. Their combined simplicity, efficiency, and precision make this novel approach particularly well suited for small samples, and may be extended to large descriptive systems or heterogeneous samples. Using a convenient sample of 20 respondents, we have demonstrated how to estimate 20 main effects using ZBT and constrained conditional logit models. We recognize that this is just one example; yet, we believe that this brief introduction to preference paths and kaizen tasks may promote future methodological research and further applications.

6 Glossary

Descriptive framework The qualitative architecture supporting a choice task, including descriptions of the context, scenarios, objects, attributes, and levels. This framework systematically conveys all aspects of the task, except for the question and choice sets, which depend on the task and its experimental design. For example, two tasks may employ the same descriptive framework but ask different questions or present different choice sets.

Descriptive system A system of attributes and levels used to express similarities and differences between objects. Such a system may provide definitions, stipulate errant combinations of attribute levels, and demarcate the proper use of adjectival statements, fonts, and graphics.

Full and partial profile A description of an object using a descriptive system. A full profile describes all attributes, and a partial profile describes a subset of attributes.

Hold-out An attribute common to all alternatives within a choice set.

Indifference statement The difference in utility between two objects is small (i.e., $|A-B| < \epsilon$). For example, reporting near equivalence between two objects is commonly used as the stopping criterion in a threshold task, such as a time trade-off.

Kaizen A Japanese term meaning “continuous improvement.”

Kaizen task An adaptive single-profile task that elicits a preference path and satisfies the conditions of the Markov property.

The Markov property The likelihood of each choice is between zero and one and depends only on the choice set, not on any prior choice.

Non-trader A person with infinitely positive or negative utility for any object with a specific attribute level, making their choices deterministic. Informally, this choice is known as a “hard no.”

Nominal attribute An attribute such that respondents disagree on the order of its levels.

Numeraire An ordinal attribute in a descriptive system that may be adjusted to achieve an indifference threshold. For example, a person may express their willingness to pay for an object in monetary terms and express indifference given a specific quantity of numeraire by stating “I don’t know.”

Opt-out An object described as a null alternative across choice tasks. For example, a Hobson’s choice is a paired comparison with a single profile and an opt-out, asking “take it or leave it?”

Ordinal attribute An attribute such that respondents agree on the order of its levels.

Pivot An object applied across choice tasks as a reference or base case. For example, a pivoted paired comparison may describe its two alternatives using a subset of attributes (i.e., partial profiles) that modify the pivot.

Preference path A sequence of two or more choices showing the evolution of an object based on an adaptive process.

Single-profile task Any task that begins by describing a single object (not multiple objects) and elicits responses regarding its profile, such as an object’s attributes or potential changes to its profile. For example, case-2 best–worst scaling introduces a single object before asking about the best and worst attributes in its profile.

Study materials The interview materials and the de-identified comma-separated-values data. Each of their 16 preference paths were recorded as an integer between one and 60, indicating the order in which the attributes were selected (see the “Interview Materials” [pdf] in the electronic supplementary materials). For example, a modal response to the last profile (55243) initially opted out (i.e., object’s profile is worse than the opt-out [W]), selected the 37th attribute order, then chose the final profile (adjusted profile is better than the opt-out [B]). Therefore, the analytical dataset (see the “Data” [csv file] in the electronic supplementary materials) has 20 rows (one for each respondent) and 16 columns (one for each task) listing their response vectors (e.g., W37B).

Threshold An alternative that terminates an adaptive task when selected. For example, a kaizen task may stop once a respondent chooses “no change” or “I don’t know.”

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40271-021-00541-z>.

Acknowledgements To produce the worked example for this paper, 18 colleagues who have previously collaborated with the co-authors were interviewed, namely (in alphabetical order): Silvia C. Craig, Sarah Dewilde, Aureliano Finch, Michał Kosma Jakubczyk, Bas Janssen, Marcel F. Jonker, Maksat Jumamyradov, Suzana Karim, Erica Lubetkin, Andrea Monteiro, Richard Norman, Mark Oppe, Jan Ostermann, Stephen Poteet, Fanni Rencz, Lucila Rey Ares, Bram Roudijk, and Elly A. Stolk. Each consented to be acknowledged for their responses, which were de-identified and posted as part of the electronic supplemental materials. Some provided further reflections during the interview and feedback on the working paper, which were greatly appreciated. We thank them for their time and support. The views, thoughts, and opinions expressed in the text belong solely to the co-authors, and not necessarily to their employer, organization, committee or other group or individual, including our esteemed colleagues.

Author Contributions BMC, KR, and JH developed the research question, generated the methods, interpreted the results, and wrote the final manuscript. BMC designed the survey instrument, conducted the interviews, and analyzed the empirical data.

Funding BMC provided all financial support for the collection of empirical data.

Code Availability The code generated and used in the analysis and the study dataset are available from the corresponding author on reasonable request.

Declarations

Conflict of Interest BMC, KR, and JH have no conflicts to disclose.

Availability of Data and Material The datasets generated and analyzed during the current study are available in the electronic supplemental materials.

References

- White DJ. Decision theory. Chicago: Aldine Pub. Co.; 1969.
- Coombs CH. A theory of data. New York: Wiley; 1964.
- Miller GA. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol Rev*. 1956;63(2):81.
- Shiffrin RM, Nosofsky RM. Seven plus or minus two: a commentary on capacity limitations. *Psychol Rev*. 1994;101(2):357–61.
- Craig BM, Rand K. Choice defines QALYs: a US valuation of the EQ-5D-5L. *Value Health*. 2018;21:S12.
- Torrance GW. Social preferences for health states: An empirical evaluation of three measurement techniques. *Socioecon Plann Sci*. 1976;10(3):129–36.
- Hanemann WM. Willingness to pay and willingness to accept: how much can they differ? *Am Econ Rev*. 1991;81(3):635–47.
- Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. *Health Serv Res*. 1973;8(3):228–45.
- Augustovski F, et al. Peruvian valuation of the EQ-5D-5L: a direct comparison of time trade-off and discrete choice experiments. *Value Health*. 2020;23(7):880–8.
- Craig BM, et al. Quality-adjusted life-years without constant proportionality. *Value Health*. 2018;21(9):1124–31.
- Craig BM, et al. Further evidence on EQ-5D-5L preference inversion: a Brazil/US collaboration. *Qual Life Res*. 2017;26(9):2489–96.
- Craig BM, Pickard AS, Rand-Hendriksen K. Do health preferences contradict ordering of EQ-5D labels? *Qual Life Res*. 2015;24(7):1759–65.
- StataCorp. Stata Stistical Software: Release 14. College Station: StataCorp LLC; 2015.
- Jakubczyk M, et al. Choice defines value: a predictive modeling competition in health preference research. *Value Health*. 2018;21(2):229–38.
- Bock RD, Jones LV. The measurement and prediction of judgment and choice. Holden-Day series in psychology. San Francisco: Holden-Day; 1968.
- David HA. The method of paired comparisons. Griffin's statistical monographs & courses. New York: Hafner Pub. Co.; 1963.
- Zermelo E., The calculations of the results of a tournament as a maximum problem in the calculus of probabilities [German]. *Mathematische Zeitschrift*. 1928;29:436–60.
- Craig BM, et al. Health preference research: an overview. *Patient Patient Cent Outcomes Res*. 2017;10(4):507–10.
- Craig BM, et al. COVID-19 health preference research: four lessons learned. *ISPOR Value Outcomes Spotlight*. 2020;6(5):1–2.
- Ramos-Goñi JM, et al. Handling data quality issues to estimate the Spanish EQ-5D-5L value set using a hybrid interval regression approach. *Value Health*. 2018;21(5):596–604.
- Ramos-Goñi JM, et al. Valuation and modeling of EQ-5D-5L health states using a hybrid approach. *Med Care*. 2017;55(7):e51–8.
- Swait J, Marley AA. Probabilistic choice (models) as a result of balancing multiple goals. *J Math Psychol*. 2013;57(1–2):1–14.
- Vass CM, Boeri M. Mobilising the next generation of stated-preference studies: the association of access device with choice behaviour and data quality. *Patient Patient Cent Outcomes Res*. 2021;14(1):55–63.
- Hartman JD, Craig BM. Does device or connection type affect health preferences in online surveys? *Patient Patient Cent Outcomes Res*. 2019;12(6):639–50.
- Craig BM. United States COVID-19 vaccination preferences (CVP): 2020 hindsight. *Patient Patient Cent Outcomes Res*. 2021;14(3):309–18.
- Aizaki H, Fogarty J. An R package and tutorial for case 2 best-worst scaling. *J Choice Modell*. 2019;32:100171.
- Craig BM, et al. US valuation of health outcomes measured using the PROMIS-29. *Value Health*. 2014;17(8):846–53.
- Chrzan K. Using partial profile choice experiments to handle large numbers of attributes. *Int J Mark Res*. 2010;52(6):827–40.