MDM
Medical Decision Making

Medical Decision Making
2023, Vol. 43(6) 667–679
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0272989X231171912
journals.sagepub.com/home/mdm
Sage

Comparison of Preferences and Data Quality between Discrete Choice Experiments Conducted in Online and Face-to-Face Respondents

Ruixuan Jiang, Eleanor Pullenayegum, James W. Shaw, Axel Mühlbacher, Todd A. Lee, Surrey Walton, Thomas Kohlmann, Richard Norman, and A. Simon Pickard

Introduction. Discrete choice experiments (DCE) are increasingly being conducted using online panels. However, the comparability of such DCE-based preferences to traditional modes of data collection (e.g., in-person) is not well established. In this study, supervised, face-to-face DCE was compared with its unsupervised, online facsimile on face validity, respondent behavior, and modeled preferences. **Methods.** Data from face-to-face and online EQ-5D-5L health state valuation studies were compared, in which each used the same experimental design and quota sampling procedure. Respondents completed 7 binary DCE tasks comparing 2 EQ-5D-5L health states presented side by side (health states A and B). Data face validity was assessed by comparing preference patterns as a function of the severity difference between 2 health states within a task. The prevalence of potentially suspicious choice patterns (i.e., all As, all Bs, and alternating As/Bs) was compared between studies. Preference data were modeled using multinomial logit regression and compared based on dimensional contribution to overall scale and importance ranking of dimension-levels. **Results.** One thousand five Online respondents and 1,099 face-to-face screened ($F2F_S$) respondents were included in the main comparison of DCE tasks. Online respondents reported more problems on all EQ-5D dimensions except for Mobility. The face validity of the data was similar between comparators. Online respondents had a greater prevalence of potentially suspicious DCE choice patterns ([Online]: 5.3% [$F2F_S$] 2.9%, P = 0.005). When modeled, the relative contribution of each EQ-5D dimension differed between modes of administration. Online respondents weighed Mobility more importantly and Anxiety/Depression less importantly. **Discussion.** Although assessments of face validity were similar between Online and $F2F_S$, modeled preferences differed. Future analyses are needed to clarify whether differences are attributable to preference or data quality variation between modes of data collection.

Keywords
EQ-5D; discrete choice experiment; face-to-face; online

Date received: September 26, 2019; accepted: March 24, 2023

Discrete choice experiments (DCEs) can be used to capture stakeholder (e.g., patient, provider, regulatory agency, etc.) preferences, benefit-risk tradeoffs, and willingness-to-pay thresholds.[1–4] DCE results can inform shared decision making, health state valuation, regulatory decisions, and insurance coverage within health care. DCEs have a strong theoretical basis, drawing from both random utility theory and Lancaster's theory.[5,6] DCE choice tasks require respondents to trade off between alternatives to choose the

Corresponding Author
Ruixuan Jiang, Center for Observational and Real-World Evidence, Merck & Co., Inc., 126 East Lincoln Avenue P.O. Box 2000 Rahway, NJ 07065, USA; (ruixuan.jiang@merck.com).

most preferred option from those presented. This method of indicating preference is typically easy for respondents to understand.[7,8]

Historically, health preference elicitation studies have been conducted in person.[9,10] Although interviewer guidance may allow real-time correction of respondent misunderstandings, face-to-face (F2F) studies also encounter challenges to validity. Most notably, respondents may be affected by social desirability bias, providing responses to demonstrate their positive characteristics.[10,11]

Recently, online panels are increasingly used to collect preference data due to time and cost efficiencies.[9,10] Due to its simplicity in the indication preferences, DCE may be better suited for online administration than preference elicitation tasks that are more complicated to explain, such as the standard gamble and time tradeoff (TTO) tasks, which prior evidence has noted are less affected by interviewer engagement.[12] However, DCEs may be cognitively burdensome, as the respondent needs to process a large volume of information presented by the alternatives in order to make a choice, even with ample time for consideration at the respondent's convenience.[13,14] Satisificing or the use of simplifying heuristics for alternative evaluation are also concerns in both online and F2F respondents.[10,15,16]

Respondents recruited in both modes of data collection are also subject to various forms of selection bias. F2F respondents must be healthy enough to speak to an interviewer or leave their homes to participate in some studies. Online panel respondents need internet access, which varies throughout segments of the population.[17] Furthermore, in order to be recruited, they must volunteer to participate in a panel to complete surveys. Finally, no matter which mode of data collection is used, respondents who consent to survey participation often differ from those who do not. Elicited preferences from different modes of data collection are therefore likely to be divergent if respondents are anticipated to be dissimilar.

DCE-based preferences from various data collection approaches may be compiled to inform health care decision making across regulatory, clinical care, and reimbursement settings. Therefore, understanding how preferences may differ by data collection approach is imperative so that these decisions are appropriately informed. A previous comparison of online and in-person binary health care DCEs found preferences to be similar between modes of data collection.[9] However, the experimental design was not intended for estimation of a value set, so the authors evaluated the effect of respondent (both quota sampled and others) and mode characteristics on choice for an alternative within a task. A comparison of DCE-derived value sets is therefore needed to provide further evidence of similarity or dissimilarity of online and F2F DCE preferences.

Within evaluation of the preference comparability between online and F2F, the assessment of DCE data quality via definitive, standalone tests may be complex. Generally recognized standards for DCE data quality do not yet exist, and unexpected choice patterns potentially indicative of poor data quality should be evaluated in the study context.[18] However, some cross-task methods to identify DCE data quality may still be harnessed to assess differences in respondent behavior and both collected and modeled preferences to evaluate the research question for the present study.[18–20] Consistent with the previous literature, an increasing preference for milder health states with increasing difference in the health state severity and low prevalence of potentially suspicious choice patterns were considered suggestive of face validity in DCE choice data.

The international protocol for valuation of the EQ-5D-5L was based on a robust program of empirical research.[21–23] Its availability has enabled standardized, high-quality data collection for country-specific studies using the TTO and DCE.[3,12,20,24] In the United States, this protocol was conducted in F2F interviews[25] and also adapted for online administration. Therefore, there was an opportunity to evaluate the comparability of online, unattended data collection with a well-understood protocol.[12,20] Supervised, F2F DCE was compared with its unsupervised, online facsimile to compare face validity, respondent behavior, and modeled preferences. Face validity was assessed using DCE choice patterns: choices relative to difference in level sum scores of alternatives in task and prevalence of potentially suspicious choice patterns.

Center for Observational and Real-World Evidence, Merck & Co., Inc, Rahway, NJ, USA (RJ); Child Health Evaluative Sciences, Hospital for Sick Children, Toronto, Canada (EP); Patient-reported Outcomes Assessment, Bristol-Myers Squibb, Princeton, NJ, USA (JWS); Duke Department of Population Health Sciences and Duke Global Health Institute, Duke University, Durham, NC, USA, Germany (AM); Department of Pharmacy Systems, Outcomes, and Policy, University of Illinois at Chicago College of Pharmacy, Chicago, IL, USA (TAL, SW, ASP); Institute for Community Medicine, Medical University Greifswald, Greifswald, Germany (TK); Curtin University School of Public Health, Perth, Australia (RN). The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: RJ, EP, JWS, TK, RN, and ASP are members of the EuroQol group, the owners of the EQ-5D. The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Bristol-Myers Squibb and EuroQol Research Foundation. JWS is an employee and stockholder of Bristol-Myers Squibb. RJ is an employee and stockholder of Merck & Co., Inc.

Methods

Measure of Health

The EQ-5D-5L is a measure of health that describes 3,125 health states using 5 dimensions of health (i.e., Mobility, Self-Care, Usual Activities, Pain/Discomfort, and Anxiety/Depression) and 5 levels of severity (i.e., no, slight, moderate, severe, and extreme problems/unable to).[26] In the standard display, these dimensions are presented in the same order, with Mobility at the top and Anxiety/Depression at the bottom of the health state. The EQ-5D-5L health states can be described as a 5-digit numeral with 1 digit for each dimension and range from no problems on any dimension (11111) to extreme problems/unable to on all 5 dimensions (55555). The measure is used in various health care decision making and measurement applications, but its most well-known application may be generation of health state values for cost-effectiveness studies.

For an approximate assessment of overall health state severity, the dimension-level responses can be summed to form a "level sum score" (LSS), which is a simple additive summary score across dimension levels, (e.g., 12312: 1 + 2 + 3 + 1 + 2 = 9).[27] An LSS can range from 5 (for 11111) to 25 (for 55555, the worst health state described by the instrument). The LSS does obscure some differences between health states, however, as multiple health states with different combinations of dimension-levels can sum to the same LSS.

Data Sources

Choice task and experimental design. Both the Online and F2F studies used binary DCE tasks, and both alternatives in each task were described by the EQ-5D-5L (Appendix I-A).[23,25] The official EQ-5D-5L valuation experimental design of 392 unique health states paired to form 196 health state pairs was used in both studies[23] (Appendix I-B).

More details of the official experimental design can be found elsewhere.[23] Briefly, 10 pairs within the experimental design were very mild, while the remaining 186 pairs were drawn from 200 health state pairs used in a series of pilot studies to minimize D-error. The 196 final pairs were divided into 28 blocks of 7 pairs each. No dominated health state pairs (i.e., where one health state was objectively milder than the other within a pair) to evaluate preference validity nor re-presentation of tasks to evaluate test-retest reliability were included in the experimental design. Each respondent was randomly assigned a block of DCE tasks and shown the tasks in random order. The tasks did not have equivalence or opt-out choices, for example, "no preference" or "prefer not to answer." Within the experimental design, the health states per pair were designated "health state A" and "health state B." For consistency, the analyses and results will refer to health states A and B, as the left/right-side display of the alternatives was also randomized at the respondent level. Therefore, the actual displayed A and B alternatives did not always match the health state labeling in the experimental design. Respondents had to complete each task before moving onto the next.

F2F EQ-5D-5L valuation study. The F2F study recruited in person/onsite, through ResearchMatch (https://www.researchmatch.org) and through advertising in online community forums.[25] Quota sampling based on age, gender, race, and ethnicity was used to recruit respondents who matched the US adult general population as a group. During each F2F survey, the interviewer and the respondent conversed one on one in a computer-assisted personal interview using the EuroQol Valuation Technology (EQ-VT), the official software platform developed by the EuroQol Group for EQ-5D-5L valuation studies.[21,28–30] For all tasks, respondents read each alternative health state and described their thought processes out loud. These actions allowed interviewers to assess misunderstandings about the health states and tasks. No practice DCE task was present in the F2F survey in adherence to the standardized experimental design.[21,23]

Two F2F comparator groups were created: 1) all DCE responses from all F2F respondents (F2F full; $F2F_F$) and 2) all DCE responses from those who understood the DCE task and considered the tasks seriously according to interviewer judgment (F2F screened; $F2F_S$). Only the $F2F_S$ cohort results are reported in the results of the main text, as this cohort provides a comparator that highlights potential advantages of having interviewers identify problematic responses. Due to the absence of interviewer judgments on preference validity, $F2F_F$ may also be a relevant comparator. However, inclusion of multiple F2F cohorts that differ by only 35 respondents disrupts the main comparison between F2F and online preferences. Therefore, $F2F_F$ results are included in Appendix II for completeness.

Online EQ-5D-5L valuation study. The Online US valuation study of the EQ-5D-5L recruited respondents from online panels exclusive, as is typical for Web-based, unsupervised studies. A recruitment e-mail was sent to online survey panels. Interested respondents opted in by clicking an e-mail link. The online quota sampling procedure matched the F2F procedure, and if the respondent's quota band was open, then they could proceed onto the survey. The survey platform hosting and respondent recruitment were completed by SurveyEngine, a company focused on choice tasks and modeling.

The online visual presentation (Appendix I-A) was a facsimile of the EQ-VT. However, with the input of researchers experienced in online valuation studies, additional features were added to the online platform to simulate the role of the interviewer (Appendix I-C). Interviewer instructions to respondents described in the interviewer guide were operationalized into the online EQ-VT in an effort to ensure task comprehension and enhance data quality. For example, if the respondent completed a DCE task in 7 s or less, the platform displayed a message to remind the respondent to carefully consider each task. Furthermore, an example DCE task comparing 11555 and 33333 was presented to the respondent in the online survey. This example task was added to demonstrate 1) tradeoffs may be necessary between dimension levels in choosing a preferred alternative and 2) respondent must choose a preferred health state even if there is no clear preference exists. These DCE considerations would have been noted by an interviewer in the F2F study. The survey platform read aloud the 2 health states in the practice task (left to right, top to bottom) using an automated, American female voice. However, the health states of the following 7 DCE tasks were not read aloud due to concerns of increasing respondent frustration, which may lead to decreased survey completion and exacerbate selection bias.

Analyses

Across all analyses, proportions were compared using chi-squared tests, and means were compared using t-tests.

Respondent characteristics. Respondent sociodemographics and other relevant characteristics, such as self-reported health, were descriptively summarized, compared between arms, and evaluated for similarity to the US general population.

DCE choice comparisons. We analyzed choice probability in the DCE tasks as a function of the difference in LSSs between alternatives in a task for face validity. The larger the LSS difference, the more likely health state A can be perceived as preferred over health state B and the more likely the respondent is to choose health state A in the DCE task. However, comparing preferences by LSS difference results in loss of detail as multiple health state pairs share the same LSS difference and is therefore imperfect in application to assessment of preference elicitation quality because respondents may not consider all dimensions to be similarly weighted/preferred. Thus, preference patterns by each health state pair were also compared between approaches.

The prevalence of DCE response patterns that may indicate task simplification and poor data validity, such as only choosing health states presented on a single side "flatlining" (i.e., all As or all Bs) and alternating left/right choices, were evaluated between arms. Of the methods put forth by the Johnson et al.[18] team, only attribute dominance, in which the respondent often chooses the health state with the better health on a single attribute regardless of other attributes, could be evaluated in these data. Time spent per task can be associated with better consideration of the alternatives,[31] but time spent reading the health states aloud in F2F surveys may bias the comparison. However, time spent per task was still compared but with particular focus on the standard deviation between comparators to provide insight on spread of time spent.

DCE modeling comparisons. The DCE-based preferences were first modeled using a multinomial logit (MNL) on a latent utility scale. The model estimated 20 regular dummy variables (Equation 1). β_n is the utility associated with each dimension-level decrement from level 1 to the dimension level associated with the dummy, (e.g., β_1 is the utility decrement from MO1 to MO2), i is the respondent, t is the choice alternative in choice sets, U_{it} represents latent utility, and ε_{it} is the residual term with an extreme value distribution.

$$\begin{aligned} U_{it} = {} & \beta_1(MO2)_{ij} + \beta_2(MO3)_{ij} + \beta_3(MO4)_{ij} + \beta_4(MO5)_{ij} \\ & + \beta_5(SC2)_{ij} + \beta_6(SC3)_{ij} + \beta_7(SC4)_{ij} + \beta_8(SC5)_{ij} + \beta_9(UA2)_{ij} \\ & + \beta_{10}(UA3)_{ij} + \beta_{11}(UA4)_{ij} + \beta_{12}(UA5)_{ij} + \beta_{13}(PD2)_{ij} \\ & + \beta_{14}(PD3)_{ij} + \beta_{15}(PD4)_{ij} + \beta_{16}(PD5)_{ij} + \beta_{17}(AD2)_{ij} \\ & + \beta_{18}(AD3)_{ij} + \beta_{19}(AD4)_{ij} + \beta_{20}(AD5)_{ij} + \varepsilon_{it} \end{aligned} \tag{1}$$

Because utility can be estimated only on a latent scale using DCE preferences, scale heterogeneity, which occurs when the probabilistic portion of the estimation model differs between comparators, can affect the values of the estimated utility weights. The Swait and Louviere test has been used to determine whether the estimated preferences differ by more than scaling factor.[32,33] In the present analyses, the Swait and Louviere test was adapted to test whether the null hypothesis (i.e., the samples had the same preferences that differ only by a scaling parameter) can be rejected by comparing log likelihoods between a model with separate parameters for both F2F and Online samples and a model with sample parameters related by a scaling parameter. A line can also be fit to the estimated preferences and associated 95% confidence intervals to visually assess whether the preferences simply differ by a multiplicative constant (i.e., scaling parameter).

Modeled dimension weights could also not be directly compared between Online and F2F as they could be estimated on only an unanchored, unitless, latent scale and were affected by the scaling parameter.[33] However, the relative dimension ranking and dimension relative importance, estimated by dividing the utility weights for a specific dimension level 5 by the sum of all dimension level 5 utility weights (e.g., MO5/[MO5 + SC5 + UA5 + PD5 + AD5]), could be compared. The number of preference inversions (i.e., utility decrement for a dimension level is less than that for its adjacent, milder dimension level) was also compared between estimated value sets.

In a binary DCE task, respondents may have a bias toward either the leftmost or rightmost option. This bias can be assessed by maintaining the left/right ordering that was presented to each respondent in the data structure and estimating an intercept and a mode-specific dummy variable (Equation 2). The intercept represented the overall tendency to choose the alternative on the left (β_0) and the mode-specific dummy variable represented the additional tendency for respondents in the Online comparator to do so (α).

$$\begin{aligned}U_{it} = {} & \beta_0 + \alpha(\text{Online}) + \beta_1(\text{MO2})_{ij} + \beta_2(\text{MO3})_{ij} + \beta_3(\text{MO4})_{ij} \\ & + \beta_4(\text{MO5})_{ij} + \beta_5(\text{SC2})_{ij} + \beta_6(\text{SC3})_{ij} + \beta_7(\text{SC4})_{ij} + \beta_8(\text{SC5})_{ij} \\ & + \beta_9(\text{UA2})_{ij} + \beta_{10}(\text{UA3})_{ij} + \beta_{11}(\text{UA4})_{ij} + \beta_{12}(\text{UA5})_{ij} \\ & + \beta_{13}(\text{PD2})_{ij} + \beta_{14}(\text{PD3})_{ij} + \beta_{15}(\text{PD4})_{ij} + \beta_{16}(\text{PD5})_{ij} \\ & + \beta_{17}(\text{AD2})_{ij} + \beta_{18}(\text{AD3})_{ij} + \beta_{19}(\text{AD4})_{ij} + \beta_{20}(\text{AD5})_{ij} + \varepsilon_{it}\end{aligned} \tag{2}$$

In exploratory analyses, data were also fit with mixed logit (MXL) models with a range of random effects to account for preference heterogeneity from various sources. Three different MXL models were fit, each including a random effect for respondents: 6 total random effects with 1 random effect per EQ-5D-5L dimension, 11 total random effects with 1 random effect for dimension levels 4 and 5, and 21 total random effects with 1 random effect for each dimension level.

Results

Respondents

A total of 1,134 F2F (F2F$_F$) and 1,005 Online respondents completed DCE tasks (Table 1). Interviewers identified 35 F2F respondents who did not understand the DCE or consider the task seriously, leaving 1,099 respondents for the F2F$_S$ sample. The F2F$_F$, F2F$_S$, and Online samples each had 7,938, 7,393, and 7,035 choice observations, respectively (Appendix I-D). Complete results for the F2F$_F$ cohort compared with the Online sample are listed in Appendix II. The rest of the results will focus on F2F$_S$.

Each sample was similar to the US general population in terms of quota-sampled characteristics (data not shown). Compared with F2Fs, Online respondents were more likely to have children younger than 18 y, be more educated, and be born in the United States (Table 1). At the group level, online respondents were also generally more ill than F2F respondents were, as demonstrated by lower mean visual analog scale values and more severe problems on Self-Care, Usual Activities, Pain/Discomfort, and Anxiety/Depression. However, Online and F2F$_S$ respondents had similar distribution of difficulty on Mobility ($P = 0.07$).

DCE Choice Patterns

Preference patterns for health states A and B appeared generally as expected (i.e., as the LSS of B lowered in contrast to A, it was more likely to be preferred; Figure 1) and similar between online and F2F$_S$. However, when the prevalence of preference for health state A was evaluated for each binary DCE task, F2F$_S$ and Online clearly had different preferences for health states (Appendix I-E).

A larger proportion of Online respondents exhibited potentially suspicious DCE choice patterns than F2F$_S$ (Table 2). Approximately 1.6% and 3.7% of Online respondents preferred health states on a single side (all As or Bs) and alternating sides, respectively. In comparison, only 0.5% and 2.2% of the F2F$_S$ respondents had the same choice patterns ($P = 0.009$ and 0.041) The prevalence of either potentially suspicious choice pattern differed between Online and F2F$_S$ ($P = 0.002$). The percentage of respondents who chose preferred health states according to a dominant attribute did not differ statistically between comparators (F2F$_s$ 43.6% v. Online 40.4%, $P = 0.140$). The standard deviation values of time spent per task for F2F$_S$ (31.0 s) and Online (34. 8s) respondents were numerically similar.

Table 1 Respondent Characteristics

Characteristic	(1) F2F$_S$ (n = 1,099)	(2) Online (n = 1,005)	P Value of (1) versus (2)
Age, y, $\bar{x}$ (s)	46.7 (18.1)	45.5 (15.2)	0.10
18–34	353 (32.1)	312 (31.0)	0.82
35–54	381 (34.7)	360 (35.8)	
55+	365 (33.2)	333 (33.1)	
Range	18–99	16–85	
Gender, n (%)			
Male	544 (49.5)	488 (48.6)	0.09
Female	550 (50.0)	517 (51.4)	
Gender, other	5 (0.5)	0 (0)	
Race, n (%)			
White	677 (61.6)	797 (79.3)	0.46
Black	144 (13.1)	121 (12.0)	
Hispanic ethnicity, n (%)	196 (17.8)	142 (14.1)	0.02
Education level greater than secondary, n (%)	716 (65.2)	710 (70.6)	<0.001
Child dependents, n (%)			
None	885 (80.6)	688 (68.5)	<0.001
Child(ren), ≤5 y old	68 (6.2)	118 (11.7)	<0.001
Child(ren), 6–17 y old	176 (16)	269 (26.8)	<0.001
Primary health insurance, n (%)			
None	93 (8.5)	103 (10.3)	0.14
Public	457 (41.6)	381 (37.9)	
Private	548 (49.9)	521 (51.8)	
Country of birth, United States	956 (87.1)	947 (94.2)	<0.001
History of illness, n (%)			
Hypertension	257 (23.4)	257 (25.6)	0.24
Arthritis	256 (23.3)	226 (22.5)	0.66
Diabetes	104 (9.5)	127 (12.6)	0.02
Heart failure	18 (1.6)	16 (1.6)	0.93
Stroke	22 (2)	20 (2.0)	0.98
Bronchitis	24 (2.2)	32 (3.2)	0.15
Asthma	130 (11.8)	98 (9.8)	0.13
Depression	285 (26)	229 (22.8)	0.09
Migraine	159 (14.5)	111 (11.0)	0.02
Cancer	64 (5.8)	20 (2.0)	<0.001
None	363 (33.1)	346 (34.4)	0.50
Health status, n (%)			
Excellent/very good/good	953 (86.8)	827 (82.3)	0.004
Fair/poor	145 (13.2)	178 (17.7)	
Self-reported EQ-VAS			
$\bar{x}$ (s)	80.4 (15.6)	73.8 (19.7)	<0.001
Median (IQR)	85 (15)	80 (23)	
Mobility			
No problems	789 (71.8)	685 (68.2)	0.07
Slight problems	202 (18.4)	187 (18.6)	
Some/moderate problems	77 (7.0)	96 (9.6)	
Severe problems	28 (2.6)	28 (2.8)	
Unable to walk about	3 (0.3)	9 (0.9)	
Self-care			
No problems	1,029 (93.6)	861 (85.7)	<0.001
Slight problems	41 (3.7)	86 (8.6)	
Some/moderate problems	25 (2.3)	43 (4.3)	
Severe problems	3 (0.3)	10 (1)	
Unable to wash or dress	1 (0.1)	5 (0.5)	
Usual activities			
No problems	827 (75.3)	666 (66.3)	0.001
Slight problems	173 (15.7)	206 (20.5)	
Some/moderate problems	79 (7.2)	103 (10.3)	
Severe problems	16 (1.5)	25 (2.5)	
Unable to do usual activities	4 (0.4)	5 (0.5)	
Pain/discomfort			
No pain or discomfort	539 (49)	368 (36.6)	<0.001
Slight pain or discomfort	363 (33)	361 (35.9)	
Moderate pain or discomfort	147 (13.4)	202 (20.1)	
Severe pain or discomfort	38 (3.5)	60 (6)	
Extreme pain or discomfort	12 (1.1)	14 (1.4)	
Anxiety/depression			
Not anxious or depressed	677 (61.6)	476 (47.3)	<0.001
Slightly anxious or depressed	264 (24.0)	268 (26.7)	
Moderately anxious or depressed	128 (11.7)	183 (18.2)	
Severely anxious or depressed	23 (2.1)	47 (4.7)	
Extremely anxious ordepressed	7 (0.6)	31 (3.1)	

F2F$_S$, face to face screened; IQR, interquartile range; VAS, visual analog scale.

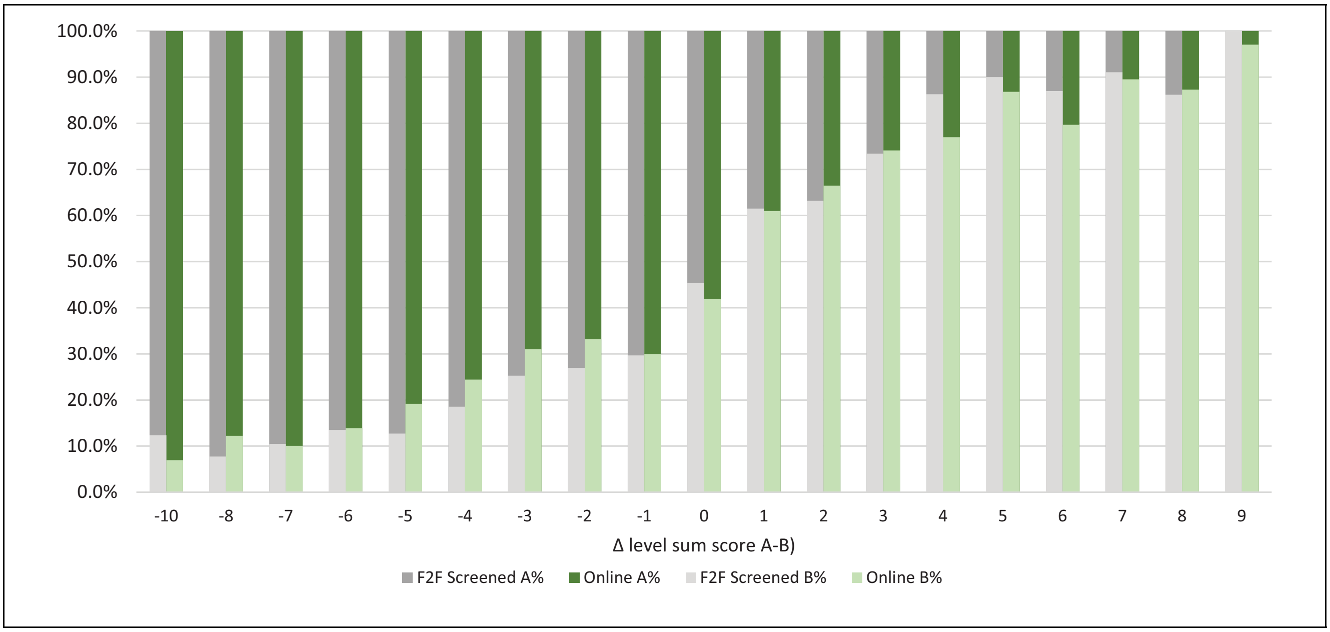

Figure 1 Relative discrete choice experiment health state preference by difference in level sum score (LSS).[a]
F2F, face-to-face. [a]The x-axis represents the difference in level sum score between health state A and health state B. When the LSS difference was negative, the LSS of health state B was greater than the LSS of health state A. Using LSS as a general measure of health state severity, health state A was expected to be preferred by a greater portion of respondents if the LSS of health state B was much greater (i.e., representing a much worse health state) than the LSS of health state A.

Table 2 Potentially Suspicious Respondent Choice Patterns and Time Spent per Task

	(1) $F2F_S$ (n = 1,099)		(2) Online (n = 1,005)		P Value, (1) versus (2)
	n	%	n	%	
Flatlining response pattern	5	0.45	16	1.59	0.009
Alternating response pattern	24	2.18	37	3.68	0.041
Either pattern	29	2.64	53	5.27	0.002
Attribute dominance	444	40.4	438	43.6	0.140
Time per task in seconds ($\bar{x}$, s)	39.3	31.0	20.5	34.8	<0.001

$F2F_S$, face to face screened. The flatlining response pattern is defined as all left-side or all right-side preferences. An alternating response pattern is defined as left/right or right/left alternating preferences.

DCE Task Comprehension

Fewer Online respondents (62.4%) strongly agreed that DCE questions were easy to understand than $F2F_S$ respondents did (69%, P = 0.025; Table 3). Interestingly, Online respondents were more certain of their DCE choices than $F2F_S$ respondents were; more Online respondents (19.9%) strongly disagreed with the statement, "I found it difficult to decide on my answers to the questions" than $F2F_S$ did (16.4%, P = 0.047). There was no statistically significant difference in comparators' self-reported ability to distinguish between alternatives with each DCE task (P = 0.448).

Modeled Preferences

The MNL-based models revealed divergent preferences between comparators. The $F2F_S$ sample had 4 insignificant parameter estimates (Table 4). The Online sample had 5 insignificant parameter estimates, 3 of which overlapped with the $F2F_S$ insignificant parameter estimates. Both main comparators had nonsignificant preference inversions, indicating that respondents did not distinguish between those 2 dimension levels, albeit at different dimension levels. MXL-based model results are listed in Appendix I-F.

Dimensional relative importance also varied by comparator in the MNL model. Mobility was more important (Online: 29% $F2F_S$: 22%), and Anxiety/Depression was less important (Online: 15% $F2F_S$: 22%) for the online comparator than $F2F_S$ (Figure 2). Similar results were reported in the exploratory MXL models (Appendix I-G). Some notable differences can also be seen in the

Table 3 Respondent Self-Reported Comprehension of DCE

DCE Comprehension Assessment	(1) F2F$_S$ (*n* = 1,099), *n* (%)	(2) Online (*n* = 1,005), *n* (%)	*P* Value, (1) versus (2)
"I found it easy to understand the questions I was asked."			
Strongly agree	758 (69.0)	627 (62.4)	0.025
Agree	302 (27.5)	329 (32.7)	
Neither agree nor disagree	27 (2.5)	36 (3.6)	
Disagree	10 (0.9)	12 (1.2)	
Strongly disagree	2 (0.2)	1 (0.1)	
"I found it difficult to decide on my answers to the questions."			
Strongly agree	151 (13.7)	145 (14.4)	0.047
Agree	328 (29.9)	246 (24.5)	
Neither agree nor disagree	160 (14.6)	156 (15.5)	
Disagree	280 (25.5)	258 (25.7)	
Strongly disagree	180 (16.4)	200 (19.9)	
"I found it easy to tell the difference between the lives I was asked to think about."			
Strongly agree	570 (51.9)	555 (55.2)	0.44
Agree	402 (36.6)	353 (35.1)	
Neither agree nor disagree	84 (7.6)	67 (6.7)	
Disagree	39 (3.6)	26 (2.6)	
Strongly disagree	4 (0.4)	4 (0.4)	

DCE, discrete choice experiment; F2F$_S$, face to face screened.

Table 4 Preference Weights by Comparator Estimated Using Multinomial Logit

	F2F$_S$ (*n* = 1,099)			Online (*n* = 1,005)		
	Estimate	SE	*P* Value	Estimate	SE	*P* Value
MO2	0.331	0.056	<0.001	0.376	0.056	<0.001
MO3	0.434[a]	0.067	<0.001	0.493[a]	0.066	<0.001
MO4	1.043	0.066	<0.001	1.132	0.066	<0.001
MO5	1.581	0.073	<0.001	1.784	0.075	<0.001
SC2	0.289	0.060	<0.001	0.197	0.060	0.001
SC3	0.261[a,b]	0.066	<0.001	0.299[a]	0.065	<0.001
SC4	0.916	0.067	<0.001	0.949	0.067	<0.001
SC5	1.211	0.065	<0.001	1.120	0.064	<0.001
UA2	0.215	0.058	<0.001	0.182	0.057	0.001
UA3	0.163[a,b]	0.064	0.011	0.352	0.064	<0.001
UA4	0.759	0.065	<0.001	0.763	0.064	<0.001
UA5	0.953	0.067	<0.001	0.892	0.065	<0.001
PD2	0.418	0.061	<0.001	0.171	0.060	0.005
PD3	0.502[a]	0.065	<0.001	0.288[a]	0.065	<0.001
PD4	1.632	0.070	<0.001	1.139	0.067	<0.001
PD5	1.903	0.072	<0.001	1.431	0.068	<0.001
AD2	0.323	0.064	<0.001	0.264	0.063	<0.001
AD3	0.564	0.064	<0.001	0.287[a]	0.064	<0.001
AD4	1.412	0.072	<0.001	0.955	0.068	<0.001
AD5	1.632	0.072	<0.001	0.915[a,b]	0.067	<0.001

AD, Anxiety/Depression; F2F$_S$, face to face screened; MO, Mobility; PD, Pain/Discomfort; SC, Self-Care; SE, standard error; UA, Usual Activities.

[a] The difference between dimension levels was not significantly different from 0.

[b] Preference inversion; none of the incremental preference inversions were significantly different from 0.

importance ranking of dimension levels between comparators based on MNL results (Table 5). Within the Online sample, Mobility level 5 was the most important, and Anxiety/Depression level 5 was 5 places lower in comparison with F2F$_S$. Other ranking differences included Pain/Discomfort level 2 being 6 steps less important and Self-Care level 3 being 4 steps more important in online compared with F2F$_S$.

The adapted Swait and Louviere log likelihood test revealed that the MNL-modeled preferences differed by more than just scale between Online and F2F$_S$ ($P < 0.001$). When plotted, the estimated preferences and associated 95% confidence intervals did not fall along a straight line, further illustrating that scaling alone did not explain the dissimilarity of modeled preferences (Figure 3). Finally, the Online comparator did not have a significant additional preference for the left-sided alternative ([F2F$_S$] $\alpha = 0.020$, $P = 0.61$).

The adapted Swait and Louviere log likelihood test showed similar results when comparing Online and F2F$_F$ ($P < 0.001$). Other corresponding comparisons of Online and F2F$_F$ for analyses can be found in Appendix II.

Discussion

Overall, the results indicated that there was comparable face validity but nuanced differences in DCE-derived preferences between data collected using a F2F approach compared with an online panel that used unattended

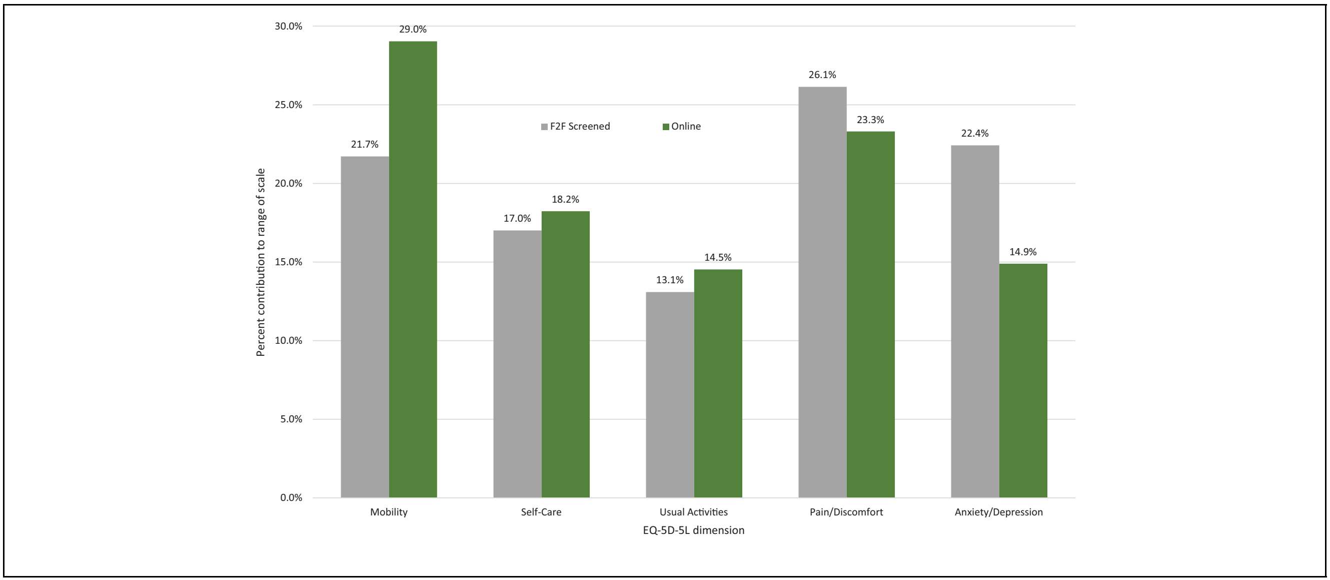

Figure 2 Relative dimension importance by mode of comparator.[a]
[a]Relative dimension importance as contribution to range of scale was estimated by dividing the utility weights for a specific dimension level 5 by the sum of all dimension level 5 utility weights (e.g., MO5/[MO5 + SC5 + UA5 + PD5 + AD5]).

Table 5 Importance Rankings of Dimension Level by Comparator[a]

		$F2F_S$	Online
Most important	1	PD5	MO5
	2	PD4	PD5
	3	AD5	PD4
	4	MO5	MO4
	5	AD4	SC5
	6	SC5	AD4
	7	MO4	SC4
	8	UA5	AD5
	9	SC4	UA5
	10	UA4	UA4
	11	AD3	MO3
	12	PD3	MO2
	13	MO3	UA3
	14	PD2	SC3
	15	MO2	PD3
	16	AD2	AD3
	17	SC2	AD2
	18	SC3	SC2
	19	UA2	UA2
Least important	20	UA3	PD2

AD, Anxiety/Depression; $F2F_S$, face to face screened; MO, Mobility; PD, Pain/Discomfort; SC, Self-Care; UA, Usual Activities. The number following denotes the severity level on a given dimension; for example, MO2 indicates level 2 on Mobility.
[a]Using multinomial logit-estimated preference weights, dimension levels within each comparator were sorted in order from most important (i.e., largest weight) to least important (i.e., smallest weight).

surveys. The overall patterns of preference for health state A or B as a function of the LSS difference were comparable (Figure 1). However, Online and F2F comparators showed divergent patterns of preferences for health state A when compared by health state pair. Online respondents provided potentially suspicious DCE choice patterns more frequently than F2F respondents did, but the overall prevalence remained low in both groups and was unlikely to have significantly affected data validity. The relative importance of dimensions differed by approach; Mobility was more important and Anxiety/Depression was less important for Online versus F2F.

Because few conclusive differences were found in the validity of DCE-based preferences between the F2F and Online comparators, it may be reasonable to conclude that both approaches were similarly valid. The study results can be evaluated in the context of other evidence to triangulate the implications of the findings. A dimension order effect may have been present in the Online comparator. Mobility was a greater portion and Anxiety/Depression was a smaller portion of the range of scale Online than in F2F. An order effect due to information-processing strategies may help explain the cause of preference discrepancies. Ryan et al.[34] evaluated respondent information processing in choice experiments by using eye tracking and found a top-to-bottom processing pattern. Without an interviewer present to ensure careful

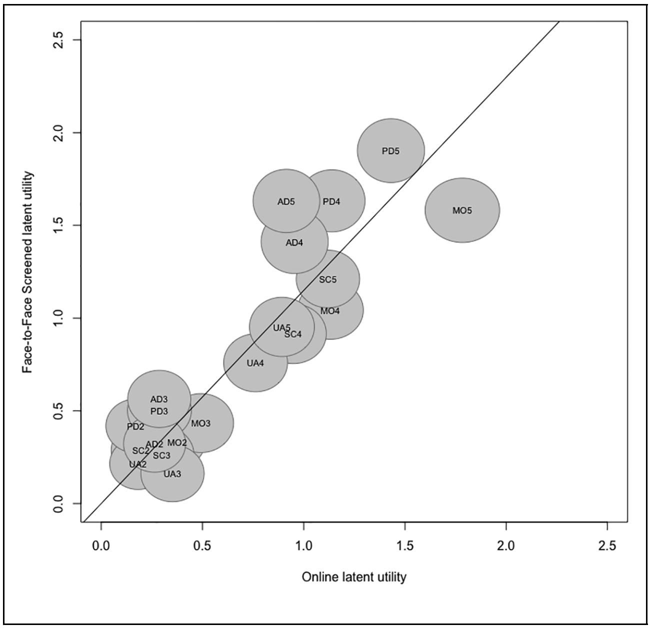

Figure 3 Relationship between dimension-level utility weights by comparator.[a]

AD, Anxiety/Depression; MO, Mobility; PD, Pain/Discomfort; SC, Self-Care; UA, Usual Activities.

[a]Multinomial logit modeled preferences and associated 95% confidence intervals (represented by circle around each dimension-level label) were plotted to determine whether the relationship between face to face screened ($F2F_S$) and Online preference weights could be interpreted as linear.

review and consideration of all dimensions, online respondents may have paid greater attention to the top few dimensions, resulting in greater importance for dimensions displayed at the top (i.e., Mobility) and lower importance for dimensions displayed at the bottom (i.e., Anxiety/Depression). Interestingly, Ryan et al. also found a left-to-right information review pattern, but the Online sample did not have significant additional preference for the left-sided alternative compared with F2F samples.

Previous studies have evaluated whether altering the order of attributes/dimensions results in different preferences with a range of results.[35–37] Mulhern et al.[35] undertook 2 separate studies to answer this research question, 1 of which found a dimension ordering effect and the other did not.[36] The null effect study may have been inadequately powered to detect preference differences using the Swait and Louviere method.[32,36] In that study, 2 alternative dimension orderings were compared with the typical display: 1) randomizing dimension levels into at the respondent level and 2) randomizing dimension levels for each task. As there were 120 possible randomized dimension orders, there was significant variability in the elicited preferences for both alternative display arms and increased difficulty to detect a difference. To reduce the cognitive burden of each task, which may also minimize the impact of dimension ordering, color coding by dimension-level severity and/or using alternatives with overlapping dimension levels have been explored with success in online panels and could be considered for implementation in future online and F2F comparisons.[38,39]

Other studies have also previously examined how preferences may differ between interviewer-guided and online, unsupervised preferences with varying results. For example, Determann et al.[10] found that no differences were found between paper and online DCEs for health insurance. Watson et al.[40] also used typical sampling frames for F2F and online survey modes and found that the willingness to pay differed by approach. Jonker et al.[41] used DCEs with EQ-5D-5L health states and online and F2F respondents as part of a broader research question and found MXL modeled responses were similar enough to be combined into a single model for analyses.

A complex range of other factors may also affect preferences for the EQ-5D-5L. The Online sample was overall more ill than the F2F sample was. While this observation underscores the ability of online sampling to recruit sicker respondents who may be unable to attend F2F interviews, health state experience has been demonstrated to affect preferences,[42] and adaptation to poorer health may help explain the divergent preference weighting of Anxiety/Depression. Online respondents reported more issues with Anxiety/Depression than F2F respondents did but weighed Anxiety/Depression less in comparison, a phenomenon sometimes seen when patients are asked to value their own health.[43] However, problems with Mobility were similarly prevalent in F2F and Online respondents, but Mobility was valued differently by respondents from the 2 approaches. Thus, adaptation to individual dimensions cannot independently explain the preference differences. During the exploratory phase of EQ-VT development, a F2F pilot study to elicit DCE-based preferences for the EQ-5D-5L in the US general population was conducted using a similar experimental design (i.e., the 200 health state pairs in the pilot study were used as the source for 186 of the health state pairs used in the present experimental design), but those preferences do not converge with the F2F results of this present study.[44,45] The pilot study found that Mobility was the most important dimension and Pain/Discomfort was the third most important dimension.[44] The $F2F_S$ sample in this comparison found the reverse: Pain/Discomfort contributed most and Mobility contributed third most to the overall scale. The preferences from the exploratory Online study and the F2F$_S$ comparators of the present study were elicited using the same mode of data collection and largely similar experimental designs, but distinctive preferences were yielded, supporting the assertion that a range of experimental, respondent, and data collection approach factors may affect estimated preferences.

The primary results for the present study were sourced from the MNL model with 20 parameters in F2Fs respondents. This base-case model could not account for preference and scale heterogeneity but was chosen as the primary practical model for several reasons. Due to each respondent only completing 7 choice tasks, MXL models may be more difficult to fit and so were undertaken as exploratory. However, the MXL-based parameter relative contributions were similar to the MNL-based results and/or MXL-based parameter standard errors were very large or 0, potentially indicative of an unstable model. These MXL model results were included in Appendix I-F for relative importance, but the MNL was used as the base-case model for the rest of the analyses. The F2F$_S$ sample was only 35 respondents fewer than the full sample, screened out for lack of DCE comprehension identified by the interviewer. Although few in comparison with the overall sample size of F2F respondents, the identification of these patients raises broader research questions regarding the application of DCE as a preference elicitation method when some in the population may be unable to express themselves through it. These questions are beyond the scope of this study but may be explored in studies with larger sample sizes or even normative discussions.

Much of the validity analyses depended on the level sum score or the sum of each dimension level within the 5-digit EQ-5D-5L health state descriptor. Because the LSS represents a score with all dimensions and dimension levels equally weighted, 2 health states cannot be distinguished if they have the same LSS. However, the LSS has shown high Pearson's and intraclass coefficient correlation values with preference-based value sets (i.e., dimension levels with unequal weighting).[46–48] Furthermore, the level sum score has been found to rank patients in terms of underlying health along a latent in nonparametric item response theory analyses.[49] We acknowledge the limitations of the LSS but maintain the LSS is useful in understanding the validity of the present study findings as a general guide to separate poorer and better health states.

There were several notable limitations to the inferences that could be drawn about the comparison of F2F and online approaches of DCE-based preference elicitation. The experimental design did not include re-presentation of a prior task nor dominated health state pairs that could have been used to make stronger conclusions regarding validity and reliability of each approach. Furthermore, this study could not separately estimate effects from mode of administration and source of respondents as they were linked in the study design by choice to reflect typical choice data collection. For example, a smaller portion of Online respondents reported DCE tasks were easy to understand; this difference may have been the result of more conscientious respondents being drawn from the Online panel and/or due to decreased social desirability bias when no interviewer was present. Personality was not measured in either comparator. Some practical choices were made in data collection using online and F2F methods, such as imposing pop-up windows if the online respondent used fewer than 7 s to answer a question and a practice DCE task. Although these design choices may have affected the comparisons, the comparators were representative of 2 preference elicitation approaches commonly applied in research. There was therefore significant value in determination of how elicited preferences diverge under normal conditions. Future research can consider a series of studies to separately estimate the separate effects of interviewer presence (attended v. unattended) and respondent source on DCE-based preferences with experimental designs from which more complex models can be estimated and inclusion of specific tasks (e.g., dominated choice task) to assess data internal validity.

The study compared preferences of F2F and online DCEs based on a rigorously tested experimental design and drawing from sampling frames typical of each mode of data collection. Although previous studies have looked at similar research questions with some of these aspects, none have used the same features to evaluate preference similarities and differences. The results point to similar face validity between preferences elicited using online and F2F approaches. However, the preferences differed between approaches. These preference differences may be due to true preference variation driven by both respondent characteristics and mode of data collection or perhaps a dimension order effect. Future studies and assessment of health care–related decisions using DCE-based preferences from multiple modes of administration should consider how data collection approach may affect the resulting preferences.

ORCID iDs

Ruixuan Jiang https://orcid.org/0000-0003-1737-2989
Eleanor Pullenayegum https://orcid.org/0000-0003-4265-1330
Richard Norman https://orcid.org/0000-0002-3112-3893

Supplemental Material

Supplementary material for this article is available on the *Medical Decision Making* website at http://journals.sagepub.com/home/mdm.

References

1. Ho MP, Gonzalez JM, Lerner HP, et al. Incorporating patient-preference evidence into regulatory decision making. *Surg Endosc*. 2015;29(10):2984–93.
2. Muhlbacher AC, Bethge S, Reed SD, Schulman KA. Patient preferences for features of health care delivery systems: a discrete choice experiment. *Health Serv Res*. 2016;51(2):704–27.
3. Ludwig K, Graf von der Schulenburg JM, Greiner W. German value set for the EQ-5D-5L. *Pharmacoeconomics*. 2018;36(6):663–74.
4. Mott DJ. Incorporating quantitative patient preference data into healthcare decision making processes: is HTA falling behind? *Patient*. 2018;11(3):249–52.
5. Vass C, Rigby D, Payne K. The role of qualitative research methods in discrete choice experiments. *Med Decis Making*. 2017;37(3):298–313.
6. Lancaster KJ. A new approach to consumer theory. *J Polit Econ*. 1966;74(2):132–57.
7. Ryan M, Gerard K. Using discrete choice experiments to value health care programmes: current practice and future research reflections. *Appl Health Econ Health Policy*. 2003;2(1):55–64.
8. Ryan M, Gerard K, Amaya-Amaya M. *Using Discrete Choice Experiments to Value Health and Health Care*. Dordrecht (the Netherlands): Springer; 2007.
9. Mulhern B, Longworth L, Brazier J, et al. Binary choice health state valuation and mode of administration: head-to-head comparison of online and CAPI. *Value Health*. 2013;16(1):104–13.
10. Determann D, Lambooij MS, Steyerberg EW, de Bekker-Grob EW, de Wit GA. Impact of survey administration mode on the results of a health-related discrete choice experiment: online and paper comparison. *Value Health*. 2017;20(7):953–60.
11. Leggett CG, Kleckner NS, Boyle KJ, Duffield JW, Mitchell RC. Social desirability bias in contingent valuation surveys administered through in-person interviews. *Land Econ*. 2003;79(4):561–75.
12. Stolk E, Ramos-Goñi JM, Ludwig K, Oppe M, Norman R.The development and strengthening of methods for valuing EQ-5D-5L—an overview. In: Devlin N, Roudijk B, Ludwig K, eds. *Value Sets for EQ-5D-5L: A Compendium, Comparative Review & User Guide*. Cham (Switzerland): Springer; 2022. p 13–27.
13. Goossens LMA, Jonker MF, Rutten-van Mölken MPMH, et al. The fold-in, fold-out design for DCE choice tasks: application to burden of disease. *Med Decis Making*. 2019;39(4):450–60.
14. Milte R, Ratcliffe J, Chen G, Lancsar E, Miller M, Crotty M. Cognitive overload? An exploration of the potential impact of cognitive functioning in discrete choice experiments with older people in health care. *Value Health*. 2014;17(5):655–9.
15. Ali S, Ronaldson S. Ordinal preference elicitation methods in health economics and health services research: using discrete choice experiments and ranking methods. *Br Med Bull*. 2012;103(1):21–44.
16. Cairns J, van der Pol M, Lloyd A. Decision making heuristics and the elicitation of preferences: being fast and frugal about the future. *Health Econ*. 2002;11(7):655–8.
17. Duffy B, Smith K, Terhanian G, Bremer J. Comparing data from online and face-to-face surveys. *Int J Mark Res*. 2005;47(6):615–39.
18. Johnson FR, Yang J-C, Reed SD. The internal validity of discrete choice experiment data: a testing tool for quantitative assessments. *Value Health*. 2019;22(2):157–60.
19. Mott DJ, Shah KK, Ramos-Goñi JM, Devlin NJ, Rivero-Arias O. Valuing EQ-5D-Y-3L health states using a discrete choice experiment: do adult and adolescent preferences differ? *Med Decis Making*. 2021;41(5):584–96.
20. Ramos-Goñi JM, Oppe M, Slaap B, Busschbach JJV, Stolk E. Quality control process for EQ-5D-5L valuation studies. *Value Health*. 2017;20(3):466–73.
21. Stolk E, Ludwig K, Rand K, van Hout B, Ramos-Goni JM. Overview, update, and lessons learned from the international EQ-5D-5L valuation work: version 2 of the EQ-5D-5L valuation protocol. *Value Health*. 2019;22(1):23–30.
22. Oppe M, Devlin NJ, van Hout B, Krabbe PF, de Charro F. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. *Value Health*. 2014;17(4):445–53.
23. Oppe M, van Hout B. The "power" of eliciting EQ-5D-5L values: the experimental design of the EQ-VT. EuroQol Working Paper Series; 2017 Available from: https://euroqol.org/wp-content/uploads/2016/10/EuroQol-Working-Paper-Series-Manuscript-17003-Mark-Oppe.pdf
24. Jensen CE, Sørensen SS, Gudex C, Jensen MB, Pedersen KM, Ehlers LH. The Danish EQ-5D-5L value set: a hybrid model using cTTO and DCE data. *Appl Health Econ Health Policy*. 2021;19(4):579–91.
25. Pickard AS, Law EH, Jiang R, et al. United States valuation of EQ-5D-5L health states using an international protocol. *Value Health*. 2019;22(8):931–41.
26. van Reenen M, Janssen B. *EQ-5D-5L User Guide-Basic Information on How to Use the EQ-5D-5 L Instrument*. Rotterdam (the Netherlands): EuroQol Group; 2013.
27. Devlin N, Parkin D, Janssen B. *Methods for Analysing and Reporting EQ-5D Data*. Cham (Switzerland): Springer; 2020.
28. Versteegh MM, Attema AE, Oppe M, Devlin NJ, Stolk EA. Time to tweak the TTO: results from a comparison of alternative specifications of the TTO. *Eur J Health Econ*. 2013;14(suppl 1):S43–51.

29. Oppe M, Rand-Hendriksen K, Shah K, Ramos-Goni JM, Luo N. EuroQol protocols for time trade-off valuation of health outcomes. *Pharmacoeconomics*. 2016;34(10):993–1004.
30. Luo N, Li M, Stolk EA, Devlin NJ. The effects of lead time and visual aids in TTO valuation: a study of the EQ-VT framework. *Eur J Health Econ*. 2013;14(suppl 1):S15–24.
31. Regier DA, Sicsic J, Watson V. Choice certainty and deliberative thinking in discrete choice experiments: a theoretical and empirical investigation. *J Econ Behav Organ*. 2019;164: 235–55.
32. Swait J, Louviere J. The role of the scale parameter in the estimation and comparison of multinomial logit models. *J Mark Res*. 1993;30(3):305–14.
33. Vass CM, Wright S, Burton M, Payne K. Scale heterogeneity in healthcare discrete choice experiments: a primer. *Patient*. 2018;11(2):167–73.
34. Ryan M, Krucien N, Hermens F. The eyes have it: using eye tracking to inform information processing strategies in multi-attributes choices. *Health Econ*. 2018;27(4):709–21.
35. Mulhern B, Shah K, Janssen MF, Longworth L, Ibbotson R. Valuing health using time trade-off and discrete choice experiment methods: does dimension order impact on health state values? *Value Health*. 2016;19(2):210–7.
36. Mulhern B, Norman R, Lorgelly P, et al. Is dimension order important when valuing health states using discrete choice experiments including duration? *Pharmacoeconomics*. 2017;35(4):439–51.
37. Norman R, Kemmler G, Viney R, et al. Order of presentation of dimensions does not systematically bias utility weights from a discrete choice experiment. *Value Health*. 2016;19(8):1033–8.
38. Jonker MF, Donkers B, de Bekker-Grob E, 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):350–63.
39. Jonker MF, Donkers B, de Bekker-Grob EW, Stolk EA. Effect of level overlap and color coding on attribute non-attendance in discrete choice experiments. *Value Health*. 2018;21(7):767–71.
40. Watson V, Porteous T, Bolt T, Ryan M. Mode and frame matter: assessing the impact of survey mode and sample frame in choice experiments. *Med Decis Making*. 2019;39(7):827–41.
41. 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):1534–47.
42. Mott DJ, Ternent L, Vale L. Do preferences differ based on respondent experience of a health issue and its treatment? A case study using a public health intervention. *Eur J Health Econ*. 2023;24(3):413–23.
43. Ubel PA, Loewenstein G, Jepson C. Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. *Qual Life Res*. 2003;12(6):599–607.
44. Krabbe PF, Devlin NJ, Stolk EA, et al. Multinational evidence of the applicability and robustness of discrete choice modeling for deriving EQ-5D-5L health-state values. *Med Care*. 2014;52(11):935–43.
45. Oppe M, van Hout B. *The "Power" of Eliciting EQ-5D-5L Values: The Experimental Design of the EQ-VT*. Rotterdam (Netherlands): EuroQol Research Foundation; 2017.
46. Lamu AN, Gamst-Klaussen T, Olsen JA. Preference weighting of health state values: what difference does it make, and why? *Value Health*. 2017;20(3):451–7.
47. Wilke CT, Pickard AS, Walton SM, Moock J, Kohlmann T, Lee TA. Statistical implications of utility weighted and equally weighted HRQL measures: an empirical study. *Health Econ*. 2010;19(1):101–10.
48. Prieto L, Sacristán JA. What is the value of social values? The uselessness of assessing health-related quality of life through preference measures. *BMC Med Res Methodol*. 2004;4:10.
49. Feng Y, Jiang R, Pickard A, Kohlmann T. Combining EQ-5D-5L items into a level summary score: demonstrating feasibility using non-parametric item response theory using an international dataset. *Qual Life Res*. 2022;31(1):11–23.