A methodological study to compare alternative modes of administration to value the EQ-5D using preference-elicitation techniques

Bryan Bennett[1], Sarah R. Hill[2], Adam Gibson[2], Yemi Oluboyede[2], Louise Longworth[2], James W. Shaw[3]

[1]Bristol Myers Squibb, United Kingdom

[2]PHMR, United Kingdom

[3]Bristol Myers Squibb, United States of America

Objectives: Two preference-elicitation techniques, the composite time trade-off (cTTO) and discrete choice experiment (DCE) without duration, are recommended to obtain EQ-5D health state utilities. These techniques can be administered using various modes of administration (MoA), but each has limitations. Face-to-face interviews are considered to produce high quality data; however, these can be costly and time consuming. Online surveys can collect large amounts of preference data quickly and cost-effectively, but evidence suggests they generate lower quality data. Remote interviewing is a novel MoA which could potentially provide a compromise between face-to-face interviews and online surveys.

The objective of this study was to explore how the MoA of preference-elicitation techniques used for valuing the EQ-5D-5L affects the quality and reliability of data. To our knowledge, this study is the first to investigate the effect of remote interviews on cTTO and DCE data quality.

Methods: Participants were allocated to two study groups, each completing 10 cTTO and 10 DCE tasks using the EQ-VT software platform. Participants conducted both an online, self-complete unassisted survey and an interviewer-assisted survey. Group A completed a face-to-face assisted MoA and Group B completed an online remote-assisted MoA. The order participants completed the unassisted and assisted surveys modes was randomised. Statistical analyses were conducted to compare the feasibility of survey completion, as well as the reliability and face validity of the data collected across the different MoA. Dichotomous outcomes were analysed using tests of proportions and continuous outcomes were analysed using t-tests. Mixed-effects regression models were fit to the data to explore the impact of MoA while accounting for participant and survey characteristics which could affect outcomes.

Results: 497 respondents completed both surveys (n=274 Group A; n=233 Group B). A lower proportion of respondents reported receiving sufficient guidance on the cTTO tasks during the unassisted survey MoA (79%) compared with the two assisted survey MoA (99-100%). However, respondents across all MoA typically reported receiving sufficient guidance on the DCE tasks. Online unassisted surveys were completed quickest ($P<0.01$) of all MoA; online remote-assisted surveys were completed quicker ($P<0.01$) than face-to-face.

Across all tests of cTTO data reliability, the unassisted surveys elicited a higher proportion ($P<0.01$) of logically inconsistent responses and a lower proportion ($P<0.01$) of responses indicating reliability of data than the two assisted surveys. There were no differences in DCE data reliability across all MoA. Differences ($P<0.01$) were observed for two tests of cTTO data reliability between the face-to-face and online remote assisted surveys; however, neither MoA was favoured in both reliability tests.

cTTO data from all MoA passed face validity checks; however, the unassisted surveys produced higher utility values for moderate and severe health-states than either assisted MoA.

Regression results confirmed that MoA was a significant predictor of observed differences in each tested outcome variable when survey and participant characteristics were accounted for.

Conclusions: Unassisted surveys are quicker to complete but yield more logically inconsistent cTTO data than assisted surveys. The reliability of DCE data is not affected by MoA. From a statistical viewpoint, both face-to-face and online remote-assisted surveys produced similar quality data.

1 Introduction

The EQ-5D is a well-established generic, preference-based instrument that was developed by the EuroQol Group in 1980 in order to measure, compare, and value health-related quality of life (HRQoL) across a wide range of disease areas (1). Due to its reliability and applicability across various disease areas, the EQ-5D has become one of the most widely used instruments to measure and value health, and it is the most frequently cited multi-attribute utility instrument, either as a preferred multi-attribute utility instrument or recommended as appropriate for use in cost-effectiveness analysis (2, 3).

Time trade-off (TTO) techniques and discrete choice experiments (DCE), without duration, are frequently used to obtain EQ-5D health state utilities (4). These techniques are recommended by the EuroQol Group and are examples of the kinds of choice-based methods recommended by the National Institute for Health and Care Excellence (NICE) for the generation of utility values.

TTO tasks require respondents to choose between 2 hypothetical lives: a shorter life in a healthy life state and a longer life in an impaired health state that is being valued (5). Different variants of TTO, such as the composite TTO (cTTO), have been developed in an attempt to mitigate potential data quality issues such as framing effects or floor effects (6). This approach has been found to yield robust results when it is administered via face-to-face interviews.

DCE techniques ask respondents to choose between sets of health states described using attributes and a range of attribute levels based on an underlying statistical design. However, a DCE that only includes the 5 EQ-5D dimensions as attributes would not produce values onto a quality-adjusted life-year scale (where 0 represents dead and 1 indicates full health) (7). A method of obtaining values from DCE on the quality-adjusted life-year scale is to combine TTO and DCE data using hybrid models, as has been done for generating some EQ-5D-5L value sets (8).

TTO and DCE can be administered using various modes of administration, but each has its own limitations. The EuroQol Group recommends face-to-face (F2F) interviews because they are considered to produce high quality data (6). However, training interviewers is vital for obtaining robust results, and conducting F2F interviews can be costly and time consuming. Online surveys are superior to F2F interviews in terms of their ability to collect large amounts of choice data quickly and cost-effectively, but they seem to generate lower quality data (9). This could occur if the tasks presented in the survey are cognitively challenging for respondents. In the absence of an interviewer who could offer guidance, respondents do not have the opportunity to access assistance to help them understand the tasks and give consistent responses (9). Remote interviewing is a novel mode of administration approach which could potentially provide a compromise between F2F interviews and online surveys (10). Existing literature examining mode of administration suggests that the mode in which stated preference tasks were administered can influence the resulting data, but the magnitude of the impact can vary depending on what methodology is used (9-15).

This study aimed to understand how the mode of administration of preference-elicitation techniques used for valuing the EQ-5D affects the quality and reliability of data. Gaining more insight into the impact the mode of administration has on responses will ensure that any future valuations of the EQ-5D will be undertaken using robust approaches that produce consistent and high-quality data.

1.1 Objectives

The aim of this research was to examine the impact of mode of administration on the conduct of cTTO and DCE methods of preference elicitation and the quality of the resulting data. This study aimed to explore whether differences are observed in responses to the same tasks between unassisted modes of administration (i.e., an online survey with no assistance from a trained interviewer) and assisted modes of administration (i.e., surveys completed with the aid of a trained interviewer either in person or remotely via video conferencing). The study also aimed to examine any differences in participants' and interviewers' perceptions of the feasibility of TTO and DCE tasks across different modes of administration.

To achieve these study aims, the 3 objectives of this research were to study the following:

- Assess the face validity and reliability of responses to cTTO and DCE tasks across different modes of administration
- Assess the feasibility of valuing health states using cTTO and DCE preference-elicitation approaches across different modes of administration
- Compare differences in feasibility, reliability, and face validity between cTTO and DCE approaches to valuing health states within each mode of administration

2 Methods

2.1 Study design

This study used cTTO and DCE tasks to value preferences for health states defined by the EQ-5D-5L. Health states were selected from the design of the EuroQol Valuation Technology (EQ-VT) protocol. One block of health states was chosen from the protocol for the cTTO and DCE tasks. The aim of this study was to evaluate any differences in responses depending on mode of administration, rather than to estimate a model to establish a full EQ-5D-5L value set; therefore, only a small sample of the EQ-VT health states were selected. The health states selected for the cTTO and DCE are discussed separately below.

Each cTTO and DCE task was developed for online and F2F administration using the computer-based EQ-VT software platform (7). Respondents were randomized into 1 of 2 groups (see Table 2.1). Each respondent completed 10 cTTO tasks and 10 DCE tasks in 2 of 3 possible modes of administration. These consisted of an unassisted online setting (UO), in which respondents were provided with a link to the survey to complete without further assistance, and 2 assisted settings: a F2F interview setting or an online setting with remote interviewer assistance (RA). In the F2F setting and RA setting, a trained interviewer guided respondents through the questionnaire. Within each study group, respondents were instructed to complete one UO survey and one of the two interviewer-assisted modes; group A completed a F2F survey and group B completed an RA survey. Each group was further divided into two study arms, in which the order that respondents completed the UO survey compared with their allocated interviewer-assisted mode survey was reversed. This approach was taken to mitigate possible ordering effects so that the resulting data can be pooled since the randomised presentation means that any bias due to order effects should not result in systematic differences between the modes. The order of survey completion each study arm is outlined in Table 2.1. Participants were instructed to complete the second survey between 4 days and 2 weeks following the first interview.

Table 2.1 Mode of administration order and group design

	Arm	Time 1	Time 2
Group A	Arm 1 (n=150)	Face-to-face	Online unassisted
	Arm 2 (n=150)	Online unassisted	Face-to-face
Group B	Arm 3 (n=150)	Online with remote assistance	Online unassisted
	Arm 4 (n=150)	Online unassisted	Online with remote assistance

Respondents completed the cTTO and DCE tasks using version 2 of the EQ-VT software (4, 7, 16). The questionnaire comprised the following sections, in order: an introduction to the study, demographic and self-reported health questions, an introduction to the cTTO tasks via 2 interactive example tasks (F2F and RA modes) or instructions (UO mode) and 3 practice tasks, 10 cTTO tasks, cTTO debriefing questions, a "feedback module" displaying the rank ordering of the health states that can be inferred from the respondent's responses to the cTTO tasks, an introduction to the DCE tasks, 10 DCE choice tasks, DCE debriefing questions, and open- and closed-ended feedback questions on the tasks they have just performed.

2.2 Health state selection

Health states for the cTTO and DCE were drawn from the set of 86 cTTO health states and 196 DCE health state pairs that constitute the cTTO and DCE designs used with the EQ-VT platform.

The health states for the cTTO in this study were drawn from block 6 of the EQ-VT cTTO experimental design (17). This block of 10 health states covers a range of severities. This study deviates from the EQ-VT design by replacing one moderate health state with a repeat state to enable a test re-test to examine reliability of TTO responses (i.e., whether a respondent has stable preferences). Table 2.2 displays all 10 health states included in the cTTO tasks. Respondents were shown the same 10 health states across all mode of administration; however, the order of health states shown to respondents was randomised.

Table 2.2 Profile, sum score, and severity of health states included in time trade-off task

Profile	Sum score	Severity
21111	6	Mild
12112	7	Mild
11212	7	Mild
23152	13	Moderate
21345	15	Moderate
34244	17	Moderate
34244	17	Moderate
55424	20	Severe
44553	21	Severe
55555	25	Severe

The health states for the DCE were drawn from block 6 of the EQ-VT DCE experimental design (17). As with the cTTO tasks, a repeated choice set was included in this study to enable a test re-test to examine the reliability of DCE choices across the mode of administration. An additional 3 choice sets were included in the DCE design for this study. These additional choice sets comprised moderate states that have been paired to make 1 health state logically dominate across the 3 choice sets, thereby enabling in-depth data quality checks of logical consistency. Table 2.3 outlines the 10 DCE choice sets included in this study. Respondents were shown the same 10 choice sets across all modes of administration; however, the order of choice sets shown to respondents was randomised and the position of each choice set on the screen was also randomised.

Table 2.3 Profile, sum score, and sum difference for each health state pair included in the discrete choice experiment task

Choice A	Sum score	Choice B	Sum score	Sum difference
51354	18	41335	16	2
24314	14	43222	13	1
12253	13	12551	14	1
13432	13	13245	15	2
13432	13	13245	15	2
43244	17	25522	16	1
23513	14	52254	18	4
32223	12	42334	16	4
32223	12	42233	14	2
42334	16	42233	14	2

2.3 Study sample

Participants were recruited from the UK public, before being sequentially assigned into one of the four study arms outlined in Table 2.1. Participants were recruited using a combination of door-to-door using social distancing and hygiene measures and online. Participants needed to have access to a laptop, desktop, or tablet computer. Respondents having access to only a smartphone were excluded because a smartphone is not suited to the completion of cTTO tasks. Participants were paid an incentive via electronic bank transfer or cheque for taking part in the study, conditional upon completing both modes of administration at the two study time points. The intended overall sample aimed to be broadly representative of the UK general population, with soft recruitment quotas applied for age group and gender.

Participants were required to meet three criteria to be eligible for participation in this study:

- To be aged 18 or over
- To be a resident of the United Kingdom and be able to read and speak English
- To have no prior experience in completing TTO or DCE tasks.

In addition to not meeting the inclusion criteria, members of the UK general population were excluded from participation if they:

- Did not have a suitable device with active internet connection to complete the TTO and DCE tasks
- Had a health condition that would affect their ability to participate or complete the study tasks (e.g., cognitive impairment)
- Self-excluded themselves from participation due to shielding guidance or health concerns in relation to the current Covid-19 global pandemic that would affect willingness to participate in F2F interviews.

2.3.1 Sample size

A total sample of n=576 (n=283 per group) was calculated for statistical power to conduct descriptive analyses and assess differences in values and proportions between mode of administration (assuming power of 0.8, significance of 0.05, standard deviation of 0.6, and an expected effect size of 0.1). As such, the study aimed to recruit a sample size of n=600 (n=300 for each group) to accommodate participants who may be unable to be included in the final data analysis (e.g., participants who fail to complete both surveys).

2.4 Statistical analysis

The analyses presented below to examine feasibility, face validity, and reliability of the study data were designed based on the EQ-VT quality check criteria (18) and data quality assessments reported in published EQ-5D valuation studies [e.g., Devlin et al. (8), Oppe et al. (4), Janssen et al. (6)]. Additionally, commonly used tests of reliability for cTTO and DCE studies were included (e.g., test re-test and dominance/transitivity tests).

Respondents in the study were allocated two identifiers: a survey identifier which was unique to every survey conducted in the study and a person identifier which was unique to each respondent. Both surveys completed by the same respondent were matched based on the person identifier to enable within-respondent analysis. Data collection ran from October 2021 to February 2022. All analyses reported below were conducted between February and March 2022 using the statistical software package Stata 15.

Primary analyses were conducted both within respondents (to compare F2F with UO surveys and to compare RA with UO surveys) and between respondents (to compare F2F and RA Mode of administration). Different analytical approaches were taken for the within-respondent and between-respondent analyses to account for paired and independent data.

2.4.1 Within-respondent analysis

Within-respondent analyses compared outcomes from online, unassisted surveys with assisted mode surveys (either F2F or RA). Each respondent included in the within-respondent analysis completed both an assisted and unassisted survey at 2 time-points. Only data for respondents who completed both their allocated surveys is included in the within-respondent analysis. The data for comparison within respondents is, therefore, not independent and required analysis suitable for paired data. To compare dichotomous outcomes (e.g., whether a response was logically consistent or logically inconsistent) a dependent-sample test of proportions (McNemar's test) was employed. For continuous outcomes data (e.g., time durations to complete tasks), paired-sample t-tests were employed. For ordinal data with more than 2 options (e.g., responses to feedback questions which are in the form of a 5-item Likert scale), Spearman's rank order correlations were conducted. Effect sizes for the Spearman's rho (ρ) were interpreted using guidance from Rea & Parker (19).

2.4.2 Between-respondent analysis

Between-respondent analyses compared outcomes between the assisted mode surveys (F2F and RA). Survey data from respondents who completed their allocated assisted-mode survey is included in the between-respondent analysis, even if those respondents failed to complete their second, unassisted survey. The data for comparison between respondents is independent. An independent-samples test of proportions was used to compare dichotomous outcomes (e.g., whether a response was logically consistent or logically inconsistent). Continuous outcomes data (e.g., time durations to complete tasks) were compared using independent-sample t-tests, and ordinal data with more than 2 options (e.g., responses to feedback questions) were compared using non-parametric Mann-Whitney rank tests.

2.4.3 Feasibility analysis

Feasibility was examined by comparing outcomes between Mode of administration on 5 elements: 1) respondent feedback following the cTTO tasks, 2) respondent feedback following the DCE tasks, 3) the mean time duration to complete the survey, 4) the number of moves to reach the point of indifference during the cTTO tasks/shortcutting the cTTO tasks, and 5) response rates by arm.

Shortcutting the cTTO tasks refers to respondents identifying their indifference point in 2 or fewer moves in the cTTO iteration sequence (4). Shortcutting suggests a lack of engagement in the task as respondents want to get through the task as quickly as possible, rather than taking the time to reach their point of indifference.

Response rates were calculated based on the proportion of individuals that agreed to take part in the study and were allocated to a study arm who then subsequently completed at least 1 of their 2 allocated surveys. Individuals who agreed to take part and were allocated to a study arm but then failed to complete either of their allocated surveys were considered as non-responses.

2.4.4 Face validity analysis

Face validity was tested by examining the correlation between cTTO utility score and sum score. Sum score is calculated as the sum of each attribute level included in a health state (e.g., the sum score for state 55555 is 25). This value is used as a proxy for severity; larger sum scores reflect more severe health states. One would expect respondents to attach a lower utility value to more severe states; therefore, a negative relationship between sum score and utility value should be observed. The face validity of the observed responses to the cTTO were tested using this assumption by plotting mean utility values derived from the cTTO for each health state against the health state's corresponding sum score. This approach has been used previously to examine face validity of EQ-5D data (Devlin 2018; Janssen et al. 2013). A visual examination of the resulting graph was used to determine the presence of the expected negative relationship and compare the mean utility values elicited via each mode of administration.

Face validity of the modelled parameter values for both preference elicitation techniques was also examined, however, these examinations are not the focus of the current manuscript. Details are available from the authors upon request.

2.4.5 Reliability analysis

Reliability was examined by comparing outcomes between Mode of administration on 9 elements: 1) all health states valued the same in cTTO, 2) the most severe health state valued no less than mild states in cTTO, 3) expressing transitivity and logical consistency of preferences in DCE (i.e., dominant choice sets selected over dominated ones), 4) test-retest for cTTO, 5) cTTO example task speeders (i.e., respondents completing the example exercise is less than 3 minutes), 6) not viewing the lead-time cTTO example exercise, 7) cTTO task speeders (i.e., respondents completing the cTTO tasks in less than 5 minutes), 8) test-retest for DCE, and 9) DCE straightliners and other repetitive patterns (i.e., respondents always selecting the left- or right-hand option in the choice set, or alternating right- and left-hand options each task).

To assess within-survey test-retest reliability of the cTTO, a 2-way mixed effects intra-class correlation coefficient (ICC) test was conducted (20). This test identified how closely correlated the utility values calculated for each health state were. A high level of correlation indicates reliability. There are no standard values for acceptable reliability using ICC and different studies have used different values to determine reliability. The values suggested in Koo & Li (2016) were adopted for this study (<0.5=poor, 0.5 to 0.75=moderate, 0.75 to 0.9=good, >0.9=excellent).

The minimum time durations set for completing the cTTO example and main tasks were taken from the EQ-VT protocol (18) to represent sufficient engagement in the tasks. Time-stamp data was collected during the survey for the duration spent on the cTTO example task during the assisted-mode interviews. UO surveys did not include an interactive example, but rather a set of instruction pages covering the same information explaining the cTTO task as in the assisted surveys. Time stamp data were also provided for time spent viewing each instruction page, which were used as a proxy for completing the cTTO example in the UO surveys.

In the F2F and RA surveys, data were collected on whether respondents viewed the lead-time cTTO example during the interactive example exercise. Since the UO surveys did not include an interactive example, the same data could not be collected. An alternative approach used was to set a minimum time threshold for viewing the relevant instruction pages as a proxy for viewing the lead-time example. It was hypothesised that respondents who took longer on the instruction pages were more likely to have read the information. A threshold of 1/3 of the median time to view the relevant instruction pages was chosen as this is a common minimum threshold considered for survey engagement in market research.

2.4.6 Regression analysis

Regression analysis allowed for an estimation of the contribution of the Mode of administration to observed variance in outcomes while accounting for other factors that may contribute to the observed variance. Therefore, all reliability tests that were shown to display statistically significant differences between Mode of administration were included in the regression analysis. Mixed-effects models were fit to the data. As mixed-effects models can accommodate paired data, this allowed all survey data to be pooled for the analysis by using the participant identifier as a random variable.

Six models were fit to the data for each dichotomous dependent variable and 3 for each continuous dependent variable. The dichotomous dependent variables were valuing all health states the same in the cTTO, valuing the worst health state no worse than mild health states, meeting the minimum time threshold during the wheelchair example, viewing the lead-time cTTO example for states worse than dead, meeting the minimum time threshold for completing the cTTO tasks, and shortcutting the cTTO tasks. The 3 continuous dependent variables were: survey duration, time taken to view the lead-time cTTO example for states worse than dead, and time taken to complete cTTO tasks. Independent variables were selected from available participant and survey characteristics data.

2.5 Study conduct

2.5.1 Ethical approval

This study did not require formal ethical approval as it was non-interventional research being conducted within a UK general population sample and participants were not recruited via the NHS. NHS REC approval was not required for this study for recruitment within the United Kingdom. The study protocol and participant-facing documents were reviewed by an independent research ethics expert under the auspices of the Association of Research Managers and Administrators.

2.5.2 Informed Consent

Participants were asked to read an information sheet and asked whether they would provide their consent to participate in the study tasks, indicating their agreement electronically. The information sheet described the study and how response data was to be used.

2.6 Demographic characteristics

Demographic data were collected on all respondents including information on gender, age, and experience of serious illness were collected. Each participant also completed the EQ-5D-5L questionnaire [including the EuroQol Visual Analogue Scale (EQ-VAS)]. Table 2.4 outlines the demographics for participants in each arm of the study.

Table 2.4 Table of participant characteristics by study arm

	Group A (n=321)[a]		Group B (n=248)[a]			
Characteristic	**n (%)**		**n (%)**		**Total**	**P[b]**
Gender						
Female	176	(55)	173	(70)	349	**0.001**
Male	143	(45)	75	(30)	218	
Other	2	(1)	0	(0)	2	
Experience of serious illness						
Personally	94	(29)	48	(19)	143	**0.007**
In family	243	(76)	174	(70)	433	0.139
In caring for others	129	(40)	88	(35)	230	0.252
Order[c]						
Unassisted online survey completed first (Arms 2 and 4)	100	(55%)	83	(45%)	183	**0.030**
Unassisted online survey completed second (Arms 1 and 3)	174	(55%)	140	(45%)	314	
Interviews by interviewer						
1	35	(11)			35	**0.000**
2	16	(5)	33	(13)	49	
3	13	(4)	11	(4)	24	
4	69	(22)	59	(24)	128	
5	49	(15)			49	
6	35	(11)	41	(17)	76	
7	13	(4)	14	(6)	27	
8	1	(0)			1	
9	9	(3)	5	(2)	14	
10	63	(20)	76	(31)	139	
11	0	(0)	2	(1)	2	
	Mean (SD)	**Median (IQR)**	**Mean (SD)**	**Median (IQR)**	**Difference in means**	**P[d]**
Age	46	44	41	40	4	**0.000**
	(14)	(23)	(13)	(18)		
EQ-VAS score	82.17	85	82.21	85.0	-0.04	0.975
	(14.93)	(12)	(14.97)	(11.5)		

Abbreviations: IQR, interquartile ratio; SD, standard deviation P, p-value at the 5% significance level

[a]Total number of participants recruited to each arm, including respondents who complete only 1 of 2 surveys, [b]χ^2 test, [c]Only respondents completing both survey modes are reported here, [d]t-test of means. In cases where 2 interviews are completed (i.e., assisted & unassisted) data on demographics is taken from the assisted interview responses; Group A, F2F & UO; Group B, RA & UO.

The mean age of respondents was 43 years (range 18-86 years) and overall self-reported health was good (mean EQ-VAS score 82.4; range 10-100). The gender distribution across respondents was 61% female, 39% male, and <1% other. Table 2.4 shows a statistically significant difference in gender distribution and in personal experience of serious illness between arms. The difference in mean age across arms is also statistically significant.

3 Results

Complete data for a minimum of 1 survey was collected from 569 respondents (n=321 Group A, n=248 Group B). Of the total 569 respondents, 497 completed both allocated surveys (n=274 Group A, F2F and UO; n=223 Group B, RA and UO).

3.1 Feasibility

3.1.1 Within-respondent analysis results

3.1.1.1 Respondent feedback to 1) cTTO tasks and 2) DCE tasks

The correlation between respondents' feedback on the cTTO tasks between the UO and interviewer-assisted surveys (either F2F or RA) is moderate to relatively strong (ρ=0.25-0.46). Most responses to the feedback questions[1] across all modes of administration were either "strongly agree" or "agree". However, a slightly larger proportion of respondents either "disagree" or "strongly disagree" with each question in UO surveys compared with either F2F or RA, except for the feedback regarding difficulty reaching indifference during the cTTO (see footnote 1 for question wording) in which the F2F mode reported the highest proportion of "disagree" responses. A smaller proportion of respondents agreed that they received sufficient guidance during the UO survey (79%) compared with the F2F (100%) and RA (99%) surveys.

Results for the feedback following the DCE tasks mirrored those following the cTTO tasks; correlations between respondent feedback in the interviewer-assisted surveys (F2F or RA) and the UO survey were moderate to relatively strong for all questions (ρ=0.28-0.47). Most responses for all feedback questions were either "strongly agree" or "agree" except for the feedback on difficulty making a choice between the two choice options in which responses were mixed across the full Likert scale from strongly agree to strongly disagree.

3.1.1.2 Time to complete the survey

The mean survey duration for the UO surveys (group A: 1481 seconds, group B: 1264 seconds) was significantly (p<0.001) shorter than either of the interviewer-assisted modes (F2F: 2453 seconds, RA: 2181 seconds). The standard deviations around the means were also much larger for the UO surveys (group A: 957 seconds, group B: 743 seconds) compared with either F2F (501 seconds) or RA (480 seconds) surveys, indicating greater variance in survey duration when the survey is unassisted.

[1] The 4 feedback questions asked following the cTTO/DCE tasks were: 1) "It was easy to understand the questions I was asked"; 2) "I found it easy to tell the difference between the lives I was asked to think about" / "I found it easy to tell the difference between choices A and B"; 3) "I found it difficult to decide on the exact points where Life A and Life B were about the same" / "I found it difficult to decide between choice A and choice B"; 4) "I received sufficient guidance to successfully complete the study tasks"

3.1.1.3 Number of moves to reach indifference during the cTTO /shortcutting

A comparison of the median number of moves to reach indifference when valuing each health state showed evidence that participants in group A reached indifference in fewer moves in 60% of the health states via the UO survey (median: 3 to 8 moves) than the F2F survey (median: 5 to 9 moves). In group B there is evidence of indifference being reached in fewer moves for 20% of the health states in the UO survey (median: 3 to 8 moves) than the RA survey (median: 5 to 8 moves).

Comparing of the proportion of participants in groups A and B who “shortcut” each health state valuation task in the cTTO by reaching indifference in fewer than 2 moves showed strong evidence (p<0.05) that a higher proportion of participants shortcut cTTO tasks in the UO surveys (group A: between 23% and 34%, group B: between 19% and 36%) compared with either F2F (between 4% and 12%) or RA (between 6% and 16%) modes.

3.1.2 Between-respondent analysis results

3.1.2.1 Respondent feedback to 1) cTTO tasks and 2) DCE tasks

A significant difference (p<0.01) was observed between the F2F and RA survey responses to feedback questions 3 and 4 following the cTTO tasks (see footnote 1 for question wording). A higher proportion of “Strongly agree” responses were offered in the RA surveys (Q3:18%, Q4:90%) compared with the F2F surveys (Q3:12%, Q4:80%). Conversely, a lower proportion of “Strongly disagree” and “disagree” responses are offered for feedback question 3 in the RA surveys (21%) compared with F2F (40%).

A significant (p<0.05) difference was observed between F2F and RA surveys survey responses to all DCE feedback questions. A higher proportion of “Strongly agree” responses were offered in the RA surveys (76%, 57%, 18%, and 88% for Qs1-4, respectively) compared with the F2F surveys (65%, 49%, 10%, and 78% for Qs1-4, respectively).

3.1.2.2 Time to complete the survey

Comparing the time to complete each of the assisted mode surveys, the mean duration of all completed F2F surveys (2444 seconds) was significantly longer than the mean duration of all completed RA surveys (2198 seconds), by approximately 4 minutes.

3.1.2.3 Number of moves to reach indifference during the cTTO /shortcutting

Respondents reach indifference quicker, and are more prone to shortcutting, during cTTO health state valuations in RA surveys compared with F2F. The mean number of moves to reach indifference during each cTTO task was slightly higher in F2F surveys (between 5.6 and 7.6 moves) compared with the RA surveys (between 5.0 and 7.0 moves). The difference in mean moves between modes was significant (p<0.05) in 70% of the health states valued. A higher proportion of respondents shortcutte the cTTO tasks in the RA surveys (between 6% and 16%) than in F2F surveys (between 4% and 12%); however, this difference was only significant for 40% of the health state valuations.

3.1.3 Response rate

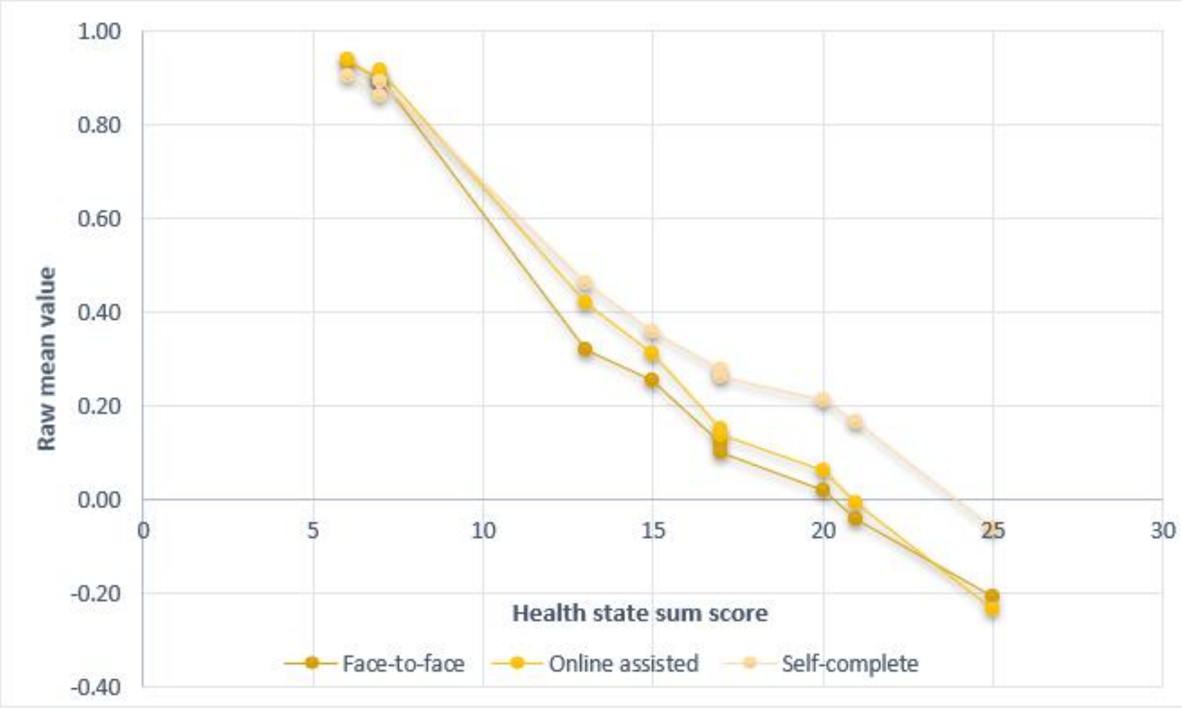
The number of respondents who were allocated to each arm and subsequently completed at least 1 of their 2 allocated surveys were examined. Within each group, response rates were compared between respondents who were allocated the UO survey first and respondents who were allocated the assisted-mode survey (F2F or RA) first. In group A, there was a significant (p<0.01) difference between the response rate of respondents who were allocated to complete the F2F survey first (96%) and the UO survey first (81%). Similarly, in group B there was also a significant (p<0.01) difference between response rates of respondents who were allocated to complete the RA survey first (86%) or the UO survey first (62%).

3.2 Face validity

Mean utility values elicited via cTTO for each health state were plotted against the sum score for each health state, representing a value of the health state's severity.

Figure 3.1 demonstrates an inverse relationship between utility value and sum score, as would be expected. Mean utility values elicited via F2F surveys are consistently lower than those elicited via the other modes, except for the worst health state (sum score 25). The largest, observable differences in mean utility values occur for the more severe states; higher utility values are elicited via the UO surveys for the 4 most severe health states compared with the F2F and RA modes.

Figure 3.1 Mean health state utility elicited via cTTO plotted against severity



3.3 Reliability

3.3.1 Within-respondent analysis results

Table 3.1 reports the results of the analyses of all within-respondent analyses with dichotomous outcomes. All analyses were conducted using dependent samples tests of proportions.

Table 3.1 Results of within-respondent reliability tests with dichotomous outcomes

Group	Test variable	Online unassisted		Online remote interviewer-assisted		F2F		McNemar's chi2	P
		N*	%*	N	%	N	%		
Valuing all health states the same in cTTO									
A	Logically inconsistent responses	20	7%	-	-	5	2%	9.0	0.004
B		16	7%	5	2%	-	-	9.3	0.003
Valuing the worst health state no less than mild states									
A	Logically inconsistent responses	70	26%	-	-	17	6%	43.2	0.000
B		42	19%	13	6%	-	-	22.7	0.000
Expressing transitivity and logical consistency of preferences in DCE									
A	Logically inconsistent responses	25	9%	-	-	17	6%	1.9	0.230
B		14	6%	9	4%	-	-	1.1	0.405
cTTO example task speeders									
A	Meeting minimum time threshold	126	46%	-	-	270	99%	140.1	0.000
B		80	36%	218	98%	-	-	132.3	0.000
Viewing lead-time cTTO example for health states worse than dead									
A	Viewing lead-time example**	230	84%	-	-	217	79%	2.3	0.165
B		171	77%	213	96%	-	-	29.4	0.000
Meeting minimum threshold time for cTTO tasks									
A	Meeting minimum time threshold	132	48%	-	-	270	99%	138.0	0.000
B		71	32%	219	98%	-	-	146.0	0.000
Test-retest for DCE									
A	Logically inconsistent responses	32	12%	-	-	31	11%	0.0	0.876
B		21	9%	17	8%	-	-	0.5	0.493
Straightlining and other repetitive patterns in the DCE									
A	Logically inconsistent responses	2	1%	-	-	3	1%	0.2	0.655
B		2	1%	3	1%	-	-	0.2	0.655

Abbreviations: P, p-value at the 5% significance level; RA, Reliability analysis

*Results are from paired test of proportions **using 1/3 of median time viewing TTO instructions as proxy for viewing lead-time example in UO surveys

3.3.1.1 Valuing all health states the same in cTTO

The proportion of logically inconsistent responses (i.e., all health states valued the same) is low across all modes of administration. However, for both groups A and B, the proportion of logically inconsistent responses is significantly (p<0.01) larger in the UO surveys compared with F2F or RA mode surveys (see Table 3.1).

3.3.1.2 Valuing the worst state no less than mild states in cTTO

The proportion of logically inconsistent responses (i.e., valuing the worst health state the same, or better than, mild states) is low for both F2F and RA mode surveys but comparably higher for the UO survey responses (see Table 3.1). The difference in proportions is significant (p<0.01) for both groups.

3.3.1.3 Expressing transitivity and logical consistency of preferences in DCE

The proportion of logically inconsistent responses (i.e., responses failing to demonstrate transitivity of preferences) is low across all Mode of administration (see Table 3.1) and there is no evidence of significant differences between UO and F2F or RA mode surveys.

3.3.1.4 cTTO example task speeders

A significantly (p<0.01) smaller proportion of respondents met the minimum time threshold (3 minutes) for viewing the cTTO example during the UO survey compared with the F2F or RA surveys (see Table 3.1). The minimum time threshold applies to the interactive example presented during F2F or RA mode surveys; no explicit threshold is set for viewing the cTTO instructions during a UO survey. The mean duration (in seconds) that respondents spent viewing the cTTO example during the UO surveys was 249 (group A) and 194 (group B) compared with the mean duration spent completing the interactive cTTO example for either the F2F (436) or RA (429) surveys. However, the comparison between the UO and F2F or RA modes was not like-for-like due to the difference in presentation of the cTTO example (see section 2.4.5 for details on presentational differences).

3.3.1.5 Viewing lead-time cTTO example for health states worse than dead

A significant (p<0.01) difference in the proportion of respondents viewing the lead-time cTTO example for health states considered worse-than-dead is observed in group B; a lower proportion viewed this example in the UO surveys compared with during the RA surveys (see Table 3.1) . There is no significant difference observed in group A between the UO and F2F surveys. As outlined in section 2.4.5, a proxy for viewing the lead-time example was used for the UO survey data; therefore, the comparison between UO and F2F or RA surveys for this outcome was not like-for-like.

3.3.1.6 Meeting minimum threshold time for cTTO tasks

Table 3.1 shows the proportion of participants meeting the minimum threshold to complete the cTTO tasks of 5 minutes is significantly (p<0.01) lower for both groups during the UO surveys compared with the F2F or RA modes. The mean duration (in seconds) to complete the 10 cTTO tasks was significantly (p<0.01) quicker in the UO surveys (group A: 348, group B: 297) compared with either F2F (672) or RA (579).

3.3.1.7 Test-retest and straightlining/repetitive patterns in the DCE

No significant differences were observed between any mode of administration in the proportion of logically inconsistent responses to either the test-retest (i.e., responses failing the DCE test re-test) or for respondents straightlining or responding in repetitive patterns during the DCE tasks. Overall, proportions of logical inconsistencies were very low for all modes of administration (see Table 3.1).

3.3.2 Between-respondent analysis results

Table 3.2 reports the results of the analyses of all between-respondent analyses with dichotomous outcomes. All analyses were conducted using independent samples tests of proportions. The between-respondent analyses which compared results between the 2 interviewer-assisted modes of administration (F2F and RA) showed no evidence of difference in the proportion of logically inconsistent outcomes between the 2 modes for any of the reliability tests (seeTable 3.2 Table 3.2).

The exception to this is the comparison of participants viewing the lead-time cTTO example for health states worse than dead. The proportion of respondents viewing the lead-time example was significantly (p<0.01) higher for the RA surveys compared with the F2F surveys.

Table 3.2 Results of between-respondent reliability tests with dichotomous outcomes

Test variable	F2F				Online remote interviewer-assisted				Difference in proportion	P
	N*	(%)	SE*	95% CI	N	(%)	SE	95% CI		
Valuing health states the same in cTTO										
Logically inconsistent responses	6	(2%)	0.008	(0.004, 0.035)	5	(2%)	0.009	(0.003, 0.039)	-0.001	0.938
Valuing worst health state no less than mild states										
Logically inconsistent responses	19	(6%)	0.014	(0.035, 0.090)	13	(5%)	0.015	(0.025, 0.083)	0.009	0.666
Expressing transitivity and logical consistency of preferences in DCE										
Logically inconsistent responses	18	(6%)	0.014	(0.033, 0.086)	12	(5%)	0.020	(0.022, 0.077)	0.01	0.626
cTTO example task speeders										
Meeting minimum threshold	299	(99%)	0.007	(0.974, 1.000)	236	(98%)	0.009	(0.961, 0.997)	0.008	0.493
Viewing lead-time cTTO example for health states worse than dead										
Viewing WTD example	239	(79%)	0.023	(0.743, 0.835)	231	(96%)	0.013	(0.933, 0.984)	-0.170	**0.000**
Meeting minimum threshold time for cTTO tasks										
Meeting minimum threshold	299	(99%)	0.007	(0.974, 1.000)	237	(98%)	0.008	(0.967, 1.000)	0.003	0.744
Test-retest for DCE										
Logically inconsistent responses	33	(11%)	0.018	(0.074, 0.144)	18	(7%)	0.017	(0.075, 0.017)	0.174	0.174
Straightlining and other repetitive patterns in the DCE										
Logically inconsistent responses	3	(1%)	0.006	(-0.001, 0.021)	3	(1%)	0.007	(-0.002, 0.026)	0.000	0.778

Abbreviations: SD, standard deviation; 95% CI, 95% confidence interval; P, p-value at the 5% significance level; *Results are from independent samples tests of proportions

3.3.2.1 Meeting minimum threshold time for cTTO tasks

Although most participants in each arm met the minimum time threshold on the cTTO main tasks for both the F2F and RA modes, a significant (p<0.01) difference in the mean duration to complete the tasks was observed. The mean time to complete the cTTO tasks during F2F surveys was 669 seconds, 89 seconds longer than the mean time for the RA surveys (590 seconds).

3.3.3 Test-retest in cTTO tasks

No statistical comparative between- or within-respondent analysis was conducted on the test re-test results for the cTTO. The ICC coefficients for each mode of administration within each group were similar (ranging from 0.874 to 0.898) and using the ICC reliability classification from Koo & Li (2016), the average correlations between the two health state valuations included in the test re-test for all modes were categorised as “good”.

3.4 Regression analysis

Nine regression models were fit in total: 3 for continuous outcome variables and 6 for dichotomous outcome variables (see section 2.4.6). In all regression models, mode of administration was a significant predictor of outcome ($p<0.01$ and $p<0.05$ for continuous and dichotomous dependent variable models, respectively). Coefficients for both F2F and RA modes were significant and favoured interviewer-assisted modes compared with the UO mode for all regression models.

Full details of the results of the regression analysis are beyond the scope of the current manuscript, however, these are available from the authors upon request.

4 Discussion

This study aimed to understand how the mode of administration of preference-elicitation techniques used for valuing the EQ-5D affects the quality and reliability of data. To achieve this, feasibility, face validity, and reliability of data elicited via DCE and cTTO methods were examined using statistical analyses. Within-respondent analyses examined paired data comparing each respondent’s responses to the UO survey and either a F2F or RA survey. Between-respondent analyses examined independent data for the interviewer-assisted mode of administration to compare data elicited from F2F surveys RA surveys directly.

4.1 Composite time trade-off data quality across modes of administration

4.1.1 Comparing unassisted online survey data with interviewer-assisted survey data

This study observed a difference in the quality of cTTO data obtained from UO surveys compared with either F2F or RA surveys. The cTTO data elicited via the UO surveys had significantly higher proportions of logically inconsistent responses (i.e., relative valuation of the worst health state compared with mild health states and valuing all health states the same) compared with each interviewer-assisted mode survey. These findings mirror existing studies comparing UO surveys with F2F interviews, that also found greater proportions of inconsistencies in responses in online surveys (12). The current study also observed evidence of lower engagement with the cTTO tasks during the UO surveys than either of the interviewer-assisted surveys. Engagement was characterised by several tests: time spent on the 10 main cTTO tasks, the proportion of participants shortcutting the valuation tasks by reaching indifference in fewer than 2 moves, time spent viewing/completing the cTTO example exercise, and viewing the lead-time cTTO example during the example exercise.

A caveat for the 2 engagement tests based on the cTTO example exercise must be made because these comparisons were not like for like between the UO and F2F or RA surveys. As outlined in Section 2.4.5, during the interviewer-assisted surveys, respondents took part in an interactive exercise to actively practice completing a cTTO. During the UO surveys, respondents did not complete a guided example exercise but viewed a series of instruction pages presenting the respondent with the same cTTO example. A proxy-measure set at one-third of the median time taken to view the relevant instruction pages was required to determine whether respondents were considered to have viewed the lead-time example during the UO surveys. However, this measure did not determine with certainty that respondents categorised as "viewing" the example absorbed the lead-time TTO example in a similar manner to respondents who went through the interactive exercise. Likewise, the examination of respondents spending sufficient time on the cTTO example was based on EuroQol's recommendation of 3 minutes. However, this was based on respondents completing an interactive exercise and no minimum threshold is recommended for respondents reading a series of instructions. Therefore, it would be expected that the time taken to view instruction pages would be less than the time taken to complete an interactive task.

Observing quicker completion of cTTO tasks in the UO surveys compared with the F2F or RA surveys was hypothesised based on existing literature that found this same pattern comparing F2F and UO surveys (Determann et al. 2017; Watson et al. 2019). The nature of interviewer-assisted surveys in which the pace of survey completion is partly dictated by how quickly the interviewer reads instructions is also likely to increase the duration of tasks. Therefore, differences in cTTO task duration would be expected between UO and F2F or RA surveys due to this factor. To determine whether time differences between UO and F2F or RA surveys represent actual differences in task engagement, the magnitude of difference between task completion times should be examined. Our findings observed the mean cTTO completion times of the UO surveys (5-6 minutes) to be almost half that of the F2F and RA surveys (10-11 minutes), a difference that was unlikely to be caused by increased interviewer-participation interaction in the F2F and RA surveys.

These conclusions regarding cTTO task engagement is supported when examining the proportion of participants "shortcutting" health state valuation tasks. "Shortcutting" may indicate reduced engagement with the tasks as participants aim to get through the tasks as quickly as possible (Janssen et al. 2013; Oppe et al. 2014). The significantly larger proportion of participants shortcutting each health state valuation in the UO surveys supports a conclusion that respondents engaged less with cTTO tasks in UO surveys compared with F2F or RA surveys.

4.1.2 Comparing face-to-face with online, remote assisted survey data

This study found insufficient evidence when examining data reliability to suggest that either F2F interviews or RA interviews provided better quality data. Each mode performed significantly worse than the other in 1 reliability test[2] and performed equally well for the remainder.

4.2 Discrete choice experiment data quality across modes of administration

In comparison with the reliability findings for cTTO data, no statistically significant differences in the proportion of logically inconsistent responses from DCE data were observed between the modes of administration in this study. This study found no evidence that UO surveys produced less reliable DCE data than F2F or RA surveys. This finding aligns with the findings of Mulhern et al (2013), who

[2] RA performed better than F2F with more respondents viewing the lead-time cTTO example yet F2F performed better than RA with respondents spending more time on the cTTO tasks

observed no effect of mode of administration on DCE responses when comparing computer-assisted personal interviews (CAPI) with UO surveys. On the other hand, Watson et al. (2019) observed contradictory findings while comparing CAPI, mail, and UO surveys, concluding that the UO survey performed better on most of their measures. However, the measures used by Watson et al. (2019) to judge mode performance focused largely on the DCE outcomes and theoretical validity rather than the measures of reliability tested in the current survey.

4.3 Face validity of data across modes of administration

The current face validity results showed significantly (p<0.05) higher mean utility values elicited for moderate to severe health states in the UO surveys compared with the F2F and RA surveys. This finding mirrors that of Norman et al (2010) who also observed higher utility values elicited via cTTO in UO surveys compared to F2F interviews. Notably, the face validity assessments were conducted only on data which passed two logical consistency checks only [i.e., not valuing all health states the same and valuing the worst health state no worse than mild states]. Therefore, factors other than a greater proportion of logical inconsistencies in UO surveys had an impact on the resulting utility values elicited via cTTO. The mean health state utilities elicited via the cTTO were extremely close between the 2 assisted modes (RA and F2F), with a difference visible in Figure 3.1 for only 1 health state (EQ-5D-5L profile: 23152), in which the F2F surveys offered a significantly (p<0.05) lower utility value than the RA surveys. This difference appeared to have been driven by a higher proportion of respondents allocating this health state the lowest utility value (-1) during the F2F surveys compared with RA, although the median utility value for this health state was the same for both modes (0.5).

Promisingly, however, data from all 3 modes of administration passed the face validity check of an inverse relationship between utility value and severity (see Figure 3.1). Therefore, although there were relative differences in the resulting utility values between UO and F2F or RA modes, the pattern of valuations aligned with prior expectations.

4.4 Practical considerations

The feedback from participants following the cTTO and DCE tasks, along with the recruitment experience, provides some insight into the practical considerations of conducting preference elicitation techniques across the range of modes of administration.

The feedback following the cTTO tasks suggests that it may be more feasible to conduct cTTO valuations via an interviewer-assisted mode than via UO surveys. Overall, most participants during all 3 modes of administration agreed with the feedback questions. However, a smaller proportion of respondents agreed that they received sufficient guidance on the cTTO tasks in the UO setting, suggesting that the cTTO instructions alone, without an interviewer to provide guidance and answer questions, may not be sufficient for many participants. Furthermore, fewer respondents in the UO surveys disagreed when asked whether they found it difficult to ‘decide on the exact points where Life A and Life B were about the same’. This suggests that a larger proportion of respondents found it difficult to reach their point of indifference in the UO arm compared with the F2F or RA arms. This could be due to greater difficulty understanding the tasks during the UO survey, which is supported by a smaller proportion of respondents agreeing that they understood the cTTO questions during the UO survey compared with the F2F or RA surveys.

The relative disparity in understanding and completing the cTTO tasks in the UO survey arm aligns with Jiang et al. (2021), who also found that fewer respondents concluded that the TTO task was easy in UO surveys than interviewer-assisted surveys.

The feedback following the DCE found little difference between responses based on mode of administration. Most respondents reported receiving sufficient guidance and reported finding the questions easy to understand. The feedback on the DCE tasks in this study differed slightly from existing literature by Rowen et al. (2016), which found that respondents completing an online DCE were more likely to report finding the wording of questions as clear than respondents in a F2F setting. This contradictory finding could be due to several factors. The DCE conducted in the study by Rowen et al. (2016) did not compare EQ-5D health states but, instead, evaluated a more complex vignette involving trade-offs in both length of life and health quality with and without treatment for a given condition. Therefore, the overall complexity of question wording is likely to be very different than when comparing EQ-5D health states. Additionally, the sample sizes of the UO group and the F2F group in the study by Rowen et al. (2016) were smaller than those used in the current study and were highly imbalanced between the UO and F2F groups, potentially impacted subsequent results.

There were significant differences identified in study response rates based on the mode of administration. Comparing between groups, response rates overall were higher in the F2F arms compared to the RA arms, suggesting a preference for F2F surveys over RA. This finding was contradictory to prior hypotheses, given the reduced time burden on respondents to attend an RA interview compared with travelling to a F2F interview. However, response rates for both groups were still reasonably high (group A: 89%, group B: 75%). This is an important practical consideration when reflecting on mode of administration for future preference-elicitation surveys. The findings from this study may be associated with the recent global coronavirus pandemic and the impact that had on in-person interaction; therefore, the relative difficulty in recruitment for the RA survey may abate over time as F2F connections become less novel and individuals suffer less from “video call fatigue” in their everyday lives.

Most interesting was comparing response rates between the order in which respondents were allocated to either an interviewer-assisted or UO survey. Across both groups, when respondents were offered an interviewer-assisted survey first response rates were high (group A: 96%, group B: 86%) compared with significantly lower response rates when respondents were offered a UO survey first (group A: 81%, group B: 62%). Examining response rates pooled to UO first compared with interviewer-assisted first (i.e., F2F or RA), a clear difference is observed; individuals were more likely to complete their first allocated survey when that survey was interviewer-assisted (91%) than UO (71%). This finding reflects that of Watson et al. (15), who observed higher response rates to CAPI interviews than online and mail surveys. The finding also aligns with feedback from the recruitment team who reported that respondents indicated they had much greater difficulty completing the UO surveys without any prior experience of the survey, compared to respondents who completed a survey with assistance first and then completed the UO survey second.

It is worth noting that because response rates were high when either a F2F or RA interview was offered first to respondents and the results from the cTTO and DCE data were very similar between the 2 modes, future studies could consider offering both F2F and RA options, allowing respondents to choose which is most convenient. In terms of survey conduct and interviewer training, the differences between the 2 modes were negligible and providing such an option may optimise results by enabling participants to complete the survey in the mode with which they are most comfortable. From an economic point of view, a hybrid-mode approach may be attractive if the cost of RA surveys is lower than that of F2F surveys because, for a given total sample size, the cost will be reduced by supplementing some F2F interviews with RA ones. However, some of the disadvantages of the RA interview can be minimised via the option of a F2F interview for those who may be unable to participate using an online format.

This survey recruited respondents using door-to-door and online recruitment approaches. Soft quotas for participant characteristics such as age and gender were applied to try and obtain a representative sample of the UK public on these characteristics, however, to maximise recruitment these were not set as hard quotas. As a result, this study offers insight into selection bias associated with these types of studies. Selection bias will be present in all preference-elicitation studies which employ an opt-in recruitment method; however, examining the demographic characteristics of both study groups provides some useful insights and considerations for future studies of these kind.

A significantly higher proportion of women took part in the study when allocated to the RA interview group (69%) compared with those allocated to the F2F interview (55%). Therefore, future studies may want to seek a more representative balance of genders, which (based on the outcomes of this study) may require more strict recruitment quotas if using only the RA mode of administration. A significant difference in mean age was also observed between the 2 groups. The mean age was higher in group A than group B. This could be a result of the technological requirements for the RA surveys in group B, with which older adults may be less familiar if they do not use video conferencing regularly (e.g., for work). Therefore, offering F2F interviews may ensure that preference-elicitation is accessible to the broadest population age range.

4.5 Further research recommendations

The current study aimed to understand the impact of different modes of administration on the feasibility and quality of stated preference surveys, and it has generated useful insights and suggestions for future researchers seeking to conduct such studies. This study has also identified further questions which are beyond its current scope to answer. To build on the research reported here, future studies may consider examining further the role that online unassisted surveys can play in eliciting stated preferences.

It is widely recognised that UO surveys are more cost-efficient to conduct compared with interviewer-assisted surveys (9). As a result of these cost efficiencies, larger sample sizes can also be recruited. However, the current study indicates that UO preference elicitation surveys may produce less reliable data (in the case of cTTO studies, DCE studies appear less affected by reliability issues). This study has also identified tests of reliability that cTTO studies perform relatively worse on than interviewer-assisted surveys. This study did not examine the relative impact on feasibility, reliability, and Face validity of data from UO surveys if respondents who were identified as offering unreliable data are removed from the data set. Further research may benefit from conducting such analysis and identifying the impact on resulting data.

Of practical importance would be identifying the proportion of respondents who perform poorly on tests of reliability, and thus the proportion of any sample that would be anticipated to be removed from final data analysis. It would be necessary for future researchers to oversample by approximately this proportion to ensure a sufficiently large, high-quality sample for analysis. Although oversampling is associated with increased survey costs, future research would be required to compare the cost of oversampling in UO surveys with the cost of the same final sample (i.e., the desired sample size for analysis, not including the oversample proportion) using interviewer-assisted survey methods.

A final consideration for future research is the possible impact on participant characteristics of removing data from the sample. One concern may be that certain participant characteristics are associated with poorer data quality; therefore, removing less reliable data may bias outcomes and result in a lack of generalisability of results. Potential solutions to this problem, if found to persist, may be to oversample from the population most at risk of providing poor data, given that not all members of this group would be expected to provide less reliable data. This would require further research to explore the effectiveness, feasibility, and cost implications of such an approach.

4.6 Strengths and limitations

This study provides novel and interesting findings which will be of interest to the research community who wish to conduct high-quality, stated preference studies. The findings from this study are based on data from a large sample of respondents and is the first (to our knowledge) to collect paired data for the direct comparison of the three modes of administration. Findings from previous studies have been based on independent samples that may be subject to differences in outcomes based on participant characteristics. The current study aimed to minimise inter-personal impacts on outcomes by requiring each respondent to complete both a UO survey and an interviewer-assisted survey at different time periods. Thus, the within-respondent analyses accounted for individual differences when comparing unassisted with assisted Mode of administration. The study design also enabled between-respondent comparisons of F2F and RA modes, akin to the studies reported in the existing literature.

Existing literature examining the impact of modes of administration to collect stated-preference data have focused on either TTO or DCE approaches. This study examined both stated preference approaches concurrently to enable a rigorous examination of both techniques within patients and subject to the same survey, interviewer, and environmental settings. The findings from this study, therefore, facilitate a deeper understanding of the relative benefits and challenges of completing different stated preference techniques under different survey conditions.

This study used the well-established EuroQol EQ-VT survey software and adhered to the most recent EQ-VT protocol (v2.0) (7) for study conduction and interviewer facilitation to ensure data collection was to a high standard. This software has been used in several successful EQ-5D validation studies globally [e.g., Jensen et al (21). Pickard et al (22), Mai et al (23)].

The analyses conducted to examine feasibility, reliability, and Face validity in this study were extensive and were selected based on the EQ-VT quality checks and existing literature in the field. In addition to individual statistical analyses for each test, regression analysis examined whether Mode of administration remained a significant predictive factor in reliability outcomes once participant and survey characteristics were accounted for. Thus, findings on mode of administration reported here can be considered reliable.

The study results reported here should also be considered with some limitations in mind. First, although several extensions to the data collection period were granted to aim to reach the specified sample size, recruitment for the group B fell short of target. Despite recruiting above sample size for the F2F arm, the final number of paired responses fell slightly short of target because of several incomplete surveys. However, the number of complete, paired responses for each arm (274 in group A and 223 in group B) are close to the 287 specified and remains larger than several other existing studies examining Mode of administration [e.g., Norman et al (2010), Mulhern et al (2013), Rowen et al (2016)]. Second, a further recruitment issue was faced during the study related to the allocation of order in which respondents completed each mode of administration. Some respondents who had been allocated to complete the UO survey first subsequently arrived at either the F2F or RA interview having failed to complete the UO survey. To minimise respondent drop-out, these participants were switched to the alternative arm to complete the UO arm second. This resulted in fewer participants completing the UO survey first than those completing it second. However, logistic regression analysis examining differences in the characteristics between respondents who completed the UO survey first and those who completed the F2F, or RA, survey first showed no differences in respondent characteristics. Therefore, the distribution of respondent characteristics in groups A and B seems not to have been affected by this reallocation of respondents.

Third, a respondent-related limitation lies in the uneven distribution of some respondent characteristics across the 2 study groups. As discussed previously, differences are observed in age, gender, and experience of personal illness. This study attempted to minimise these differences in characteristics by randomising group allocation; however, as discussed previously, response rates once respondents were randomised to a study arm were substantially lower in group B (UO and RA) than group A (UO and F2F). Furthermore, data were not collected on respondents’ education level, income, or occupation. The study is unable to comment on whether these factors contribute to differences in observed outcomes; these data should be collected for examination in future studies.

5 Conclusions

This study aimed to understand how the mode of administration of preference-elicitation techniques used for valuing the EQ-5D affects the quality and reliability of data. The findings from this study suggest that data quality from DCE studies is not impacted by mode of administration. Therefore, based on the evidence from this study, researchers conducting DCE studies can use either UO or interviewer-assisted mode of administration without concerns over data quality provided appropriately detailed instructions are provided to respondents.

On the other hand, the quality of cTTO data differed substantially between UO and F2F or RA modes of administration. Data from the UO surveys performed worse on most reliability tests and feedback following the cTTO tasks suggests that respondents' understanding of the cTTO tasks was lower in UO. Respondents also reported feeling less sufficiently guided through the tasks when instructions were shown without an interviewer to talk through the task and answer questions. The cTTO data from all modes demonstrated Face validity, but mean utility values derived from the UO surveys were consistently higher than the F2F or RA surveys for moderate to severe health states.

Based on these findings, an interviewer-assisted mode of administration is recommended for cTTO studies. However, future research has been suggested to explore whether using UO surveys to collect cTTO data may still allow high-quality data to be collected and analysed in stated preference studies.

Comparing F2F and RA modes of administration provided limited evidence to suggest either mode produces higher quality data than the other. Practical considerations should guide future researchers' choice of Mode of administration between F2F and RA surveys. However, a hybrid approach allowing respondents to select their preferred interviewer-assisted mode may have both economic and sample selection advantages.

6 References

1. Devin NJ, Brooks R. EQ-5D and the EuroQol Group: past, present, and future. Appl Health Econ Health Policy. 2017;15(2):127-37.

2. Buchholz I, Janssen MF, Kohlmann T, Feng Y-S. A systematic review of studies comparing the measurement properties of the three-level and five-level versions of the EQ-5D. PharmacoEconomics. 2018;36(6):645-61.

3. Kennedy-Martin M, Slaap B, Herdman M, van Reenen M, Kennedy-Martin T, Greiner W, et al. Which multi-attribute utility instruments are recommended for use in cost-utility analysis? A review of national health technology assessment (HTA) guidelines. Eur J Health Econ. 2020;21(8):1245-57.

4. 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.

5. Oppe M, Rand-Hendriksen K, Shah K, Ramos-Goñi JM, Luo N. EuroQol protocols for time trade-off valuation of health outcomes. PharmacoEconomics. 2016;34:993-1004.

6. Janssen BMF, Oppe M, Versteegh MM, Stolk EA. Introducing the composite time trade-off: a test of feasibility and face validity. Eur J Health Econ. 2013;14(Suppl 1):S5-S13.

7. Stolk E, Ludwig K, Rand K, van Hout B, Ramos-Goñi 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.

8. Devlin NJ, Shah K, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: an EQ-5-D-5L value set for England. Health Econ. 2018;27(1):7-22.

9. Norman R, King MT, Clarke D, Viney R, Cronin P, Street D. Does mode of administration matter? Comparison of online and face-to-face administration of a time trade-off task. Qual Life Res. 2010;19(4):499-508.

10. Shah KK, Lloyd A, Oppe M, Devin NJ. One-to-one versus group setting for conducting computer-assisted TTO studies: findings from pilot studies in England and the Netherlands. Eur J Health Econ. 2013;14(Suppl 1):S65-S73.

11. 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.

12. Jiang R, Shaw J, Mühlbacher A, Lee TA, Walton S, Kohlmann T, et al. Comparison of online and face-to-face valuation of the EQ-5D-5L using composite time trade-off. Qual Life Res. 2021;30(5):1433-44.

13. Mulhern B, Longworth L, Brazier J, Rowen D, Bansback N, Devin N, 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.

14. Rowen D, Brazier J, Keetharuth A, Tsuchiya A, Mukuria C. Comparison of modes of administration and alternative formats for eliciting societal preferences for burden of illness. Appl Health Econ Health Policy. 2016;14(1):89-104.

15. 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.

16. Yang Z, Luo N, Oppe M, Bonsel G, Busschbach J, Stolk E. Toward a smaller design for ED-5D-5L valuation studies. Value Health. 2019;22(11):1295-302.

17. Oppe M, van Hout B. The "power" of eliciting EQ-5D-5L values: the experimental design of the EQ-VT. Rotterdam, The Netherlands: EuroQol Research Foundation; 2017. Report No.: 17003.

18. Ramos-Goñi JM, Oppe M, Slaap B, Busschbach JJ, Stolk E. Quality control process for EQ-5D-5L valuation studies. Value Health. 2017;20(3):466-73.

19. Rea LM, Parker RA. Designing and Conducting Survey Research: A Comprehensive Guide. San Francisco, CA: Jossey-Bass; 1992.

20. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine. 2016;15(2):155-63.

21. 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.

22. Pickard AS, Law EH, Jiang R, Pullenayegum E, Shaw JW, Xie F, et al. United States valuation of EQ-5D-5L health states using an international protocol. Value Health. 2019;22(8):931-41.

23. Mai VQ, Sun S, Minh HV, Luo N, Giang KB, Lindholm L, et al. An EQ-5D-5L value set for Vietnam. Qual Life Res. 2020;29(7):1923-33.