A methodological study to compare alternative modes of administration for undertaking preference-elicitation studies

MSR96

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

[1]PHMR, United Kingdom; [2]Bristol Myers Squibb, United Kingdom; [3]Bristol Myers Squibb, United States of America

*National Institute for Health and Care Excellence, United Kingdom

Background

- The EQ-5D is a well-established generic, preference-based instrument developed by the EuroQol Group to measure, compare, and value health-related quality of life across a wide range of disease areas.[1]
- Preference elicitation methods, such as the time trade-off (TTO) and discrete choice experiments (DCEs), are frequently used to obtain EQ-5D health state utilities.[2]
- 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.[3]
- DCE techniques ask respondents to choose between sets of health states described using a range of attributes and levels based on an underlying statistical design.
- TTO and DCE tasks can be administered using various modes of administration, but each has its own limitations:
  - **Face-to-face (F2F)** interviews are considered to produce high quality data[4] but training interviewers is vital for obtaining robust results, and conducting F2F interviews can be costly and time consuming.
  - **Unassisted online surveys (UO)** can collect large amounts of data quickly and cost-effectively, but they seem to generate lower quality data[5] in the absence of an interviewer who can offer guidance for cognitively challenging tasks.
  - **Remote assisted interviewing (RA)** is a novel approach which could potentially provide a compromise between F2F interviews and UO surveys[6]
- Previous studies have found differences in data quality between combinations of F2F, RA, and UO surveys for either TTO or DCE approaches[5-12], but none to our knowledge have compared all 3 methods on both TTO and DCE tasks in a single study to provide a comprehensive understanding of the impact of all modes of administration.

Study Aim

This study aimed to understand how the mode of administration (MoA) used for valuing the EQ-5D with composite TTO (cTTO) and DCE methods affects the quality and reliability of data. The study aimed to explore whether differences are observed in responses to the same tasks between unassisted (UO) and assisted (F2F or RA) MoA and participants' perceptions of feasibility of completing tasks using each mode.

Objectives

- To achieve the study aim, 3 objectives of this research were to study the following:
  1. Assess the face validity and reliability of respondent responses to cTTO and DCE tasks across different modes of administration.
  2. Assess the feasibility of valuing health states using cTTO and DCE preference-elicitation approaches across different modes of administration.
  3. Compare differences in feasibility, reliability, and face validity between cTTO and DCE approaches to valuing health states within each mode of administration.

Methods

Study Design

- cTTO and DCE tasks were used to value preferences for health states defined by the EQ-5D-5L using version 2 of the EuroQol Valuation Technology (EQ-VT) software platform.[2,13,14]
- Respondents completed the survey (consisting of 10 cTTO tasks and 10 DCE tasks) twice using 2 different modes: 1 assisted (either F2F or RA) and one unassisted (UO). An interval of 4-14 days was implemented between the completion of the 2 modes of administration.
- Participants were randomised into two groups:
  - **Group A** completed a F2F interview and UO survey
  - **Group B** completed a RA interview and UO survey
- Within each group, participants were further randomised by the order in which they completed the UO survey compared with their allocated assisted MoA.
- The survey also included: demographic and self-reported health questions, example and warm-up tasks for the cTTO, and debriefing questions on both the cTTO and DCE tasks.

Health state selection

- Health states for the cTTO and DCE were drawn from the cTTO and DCE designs used for the EQ-VT platform.[15]
- Health states ranged from mild severity (e.g., EQ-5D-5L profile 21111) to severe (e.g., EQ-5D-5L profile 55555) and included similar distributions of mild, moderate, and severe health states.
- The block of health states used for the cTTO tasks were adapted from the EQ-VT design block to include 1 repeated, moderate health state (EQ-5D-5L profile 34244) to enable a test re-test to examine reliability of TTO responses across MoA.
- A repeated choice set was also included in the DCE tasks to enable a repeated assessment to evaluate test-retest reliability of DCE choices across each MoA (**Table 1**)
- Three choice sets comprising moderate states that were paired such that 1 health state logically dominated another were included in the DCE tasks to enable in-depth data quality checks of logical consistency (**Table 1**).
- Respondents were shown the same 10 cTTO health states and the same 10 DCE choice pairs in all modes of administration but the order they were presented was randomised.

Table 1. Profile and sum score for each health state pair included in the discrete choice experiment task

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

†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).

Study sample

- The intended overall sample aimed to be broadly representative of the UK general population, with soft recruitment quotas applied for age group and gender.
- Respondents were over 18, had to have access to a laptop, computer, or tablet device, resided in the UK, and have no previous experience of TTO or DCE tasks.
- Participants were recruited using a combination of door-to-door and online approaches.

Statistical analysis

- Analysis was conducted to test the **feasibility**, **face-validity,** and **reliability** of the data elicited from both the cTTO and DCE tasks across all modes (**Figure 1**).
- Data from surveys completed by the same respondent were compared to examine differences in outcomes between interviewer-assisted and unassisted modes (i.e., *within-respondent analysis of paired data*).
- Data elicited from the two assisted modes (F2F and RA) were compared across the 2 groups of respondents (Groups A and B) to examine differences in outcomes between the modes of administration (i.e., *between-respondent analysis of independent data*).
- *Within-respondent analysis:* A dependent samples test of proportions (i.e., McNemar's test) was conducted to compare dichotomous outcomes across modes and paired samples t-tests were conducted for continuous data. Spearman's rank order correlations were conducted to compare ordinal data with more than 2 options.

Statistical analysis continued

- *Between-respondent analysis:* An independent-samples test of proportions was conducted to compare dichotomous outcomes across the two assisted modes (F2F and RA) and independent samples t-tests were conducted for continuous data. Non-parametric Mann-Whitney rank tests were conducted to compare ordinal data with more than 2 options.

Figure 1. Tests of reliability, face-validity, and feasibility

Reliability tests: outcomes compared between MoA on the following elements:

1. Respondent feedback following the cTTO tasks
2. Respondent feedback following the DCE tasks
3. Mean time duration to complete the survey
4. Number of moves to reach the point of indifference during the cTTO tasks
5. Response rates

Testing face-validity

- The correlation between cTTO utility score and sum score was examined. Sum score is used as a proxy for severity; larger sum scores reflect more severe health states
- Lower utility values would be expected for more severe states; therefore, a negative relationship between sum score and utility value should be observed.

Reliability tests: outcomes compared between MoA on the following 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
4. Test-retest for cTTO
5. Speeding through the cTTO example task (i.e., completing in less than 3 minutes)
6. Not viewing the lead-time cTTO example exercise
7. Speeding through the cTTO main tasks (i.e., completing in less than 5 minutes)
8. Test-retest for DCE
9. DCE straight-liners and other repetitive response patterns to DCE questions

Participant characteristics

- 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, n=223 Group B).
- Statistically significant differences were observed in gender distribution, in personal experience of serious illness between arms, and mean age across groups **(Table 2).**

Table 2. Participant characteristics by study group

Characteristic	Group A (n=321)[a] n (%)	Group B (n=248)[a] 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]				
UO survey completed first	100 (55%)	83 (45%)	183	**0.030**
UO survey completed second	174 (55%)	140 (45%)	314	
	Mean (SD)	**Median (IQR)**	**Mean (SD)**	**Median (IQR)**	**Difference in means**	**P[d]**
Age	46 (14)	44 (23)	41 (13)	40 (18)	4	**<0.001**
EQ-VAS score	82.17 (14.93)	85 (12)	82.21 (14.97)	85.0 (11.5)	-0.04	0.975

Abbreviations: IQR, interquartile ratio; SD, standard deviation P, p-value at the 5% significance level
[a] Total number of participants recruited to each group, including respondents who complete only 1 of 2 surveys, [b] x^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.

Feasibility findings

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

- The correlation between respondents' feedback on both the cTTO and DCE tasks between the UO and interviewer-assisted surveys (either F2F or RA) was moderate to relatively strong.
- A smaller proportion of respondents agreed that they received sufficient guidance on the cTTO tasks during the UO survey (79%) compared with the F2F (100%) and RA (99%) surveys.
- More respondents disagreed that "the questions were easy to understand" in the UO surveys compared with F2F and RA surveys.

Number of moves to reach indifference during the cTTO

- A higher proportion of participants shortcut the cTTO tasks by reaching indifference in fewer than 2 moves in the UO surveys (between 19% and 36%) compared with either F2F (between 4% and 12%) or RA (between 6% and 16%) modes ($p<0.01$).

Response rate

- Response rate was examined on the number of respondents allocated to each group who completed at least 1 of the 2 surveys they were allocated to. Response rates were compared between respondents allocated to the UO survey first and those allocated to an assisted-mode survey (F2F or RA) first.
- Response rates were higher ($p<0.01$) for respondents allocated to either F2F or RA than an UO survey. Response rates were high for both assisted-mode groups (F2F: 89%; RA: 79%).

Time to complete the survey

- 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).
- Mean duration of all completed F2F surveys (2444 seconds) was significantly longer than the mean duration of all completed RA surveys (2198 seconds), by ~4 minutes.

Reliability findings

Between-respondent analysis

- The only test of logical consistency where a significant difference was observed between RA and F2F modes was the comparison of respondents **viewing the lead-time cTTO example for health states worse than dead**. The proportion of respondents viewing it was higher ($p<0.01$) for the RA surveys (96%) compared with the F2F (79%) surveys.
- No differences were observed between RA and F2F surveys in the DCE data.

Within-respondent analyses of DCE data

- **No statistically significant differences** between UO survey responses and the interviewer-assisted survey responses were observed for the 3 reliability tests of the DCE data.
- The proportion of respondents failing to demonstrate **transitivity and logical consistency of preferences** was **low** across all modes of administration (between 4% and 9%).
- The proportion of respondents failing the **test-retest task** was **similar** across all modes of administration (between 8% and 12%).
- A **negligible proportion** of respondents (1%) **straight-lined the DCE or responded in predictive patterns** during any of the modes of administration.

Within-respondent analysis of cTTO data

Statistically significant differences were observed between the UO survey responses and the interviewer-assisted survey responses for 5 of 6 reliability tests of the cTTO data.

- A higher proportion of respondents demonstrated logical inconsistencies in the UO surveys compared with both assisted modes (F2F and RA) for the following tests of feasibility:
  - Across all modes the proportion of respondents **valuing all health states the same in the cTTO** was low; however, significantly ($p<0.01$) more respondents did so in the UO mode (7%) than the F2F or RA modes (2%).
  - Across all modes the proportion of respondents **valuing the worst state no less than mild states in cTTO** was low; however, significantly ($p<0.01$) more respondents did so in the UO mode (19%-26%) than the F2F or RA modes (6% both modes).
- A higher proportion of respondents failed to meet minimum thresholds in the UO surveys compared with both assisted modes (F2F and RA) for the following tests of reliability:

The proportion of respondents meeting **the minimum threshold to complete the cTTO tasks** was lower ($p<0.01$) during the UO surveys (32%-48%) compared with the F2F (99%) or RA (98%) modes.

The **mean duration** (in seconds) was quicker ($p<0.01$) during UO surveys (Group A: 348, Group B: 297) compared with F2F (672) or RA (579) modes.

A smaller ($p<0.01$) proportion of respondents met the **minimum time threshold for viewing the cTTO example** during the UO survey (36%-46%) compared with the F2F (99%) or RA (98%) surveys.

The **mean duration** (in seconds) during UO surveys was 249 (Group A) and 194 (Group B) compared with either the F2F (436) or RA (429) surveys.

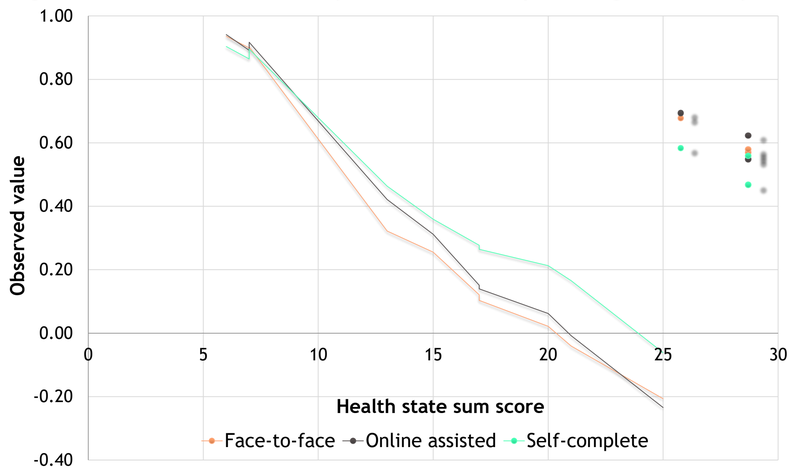
A lower ($p<0.01$) proportion of respondents **viewed the lead-time cTTO example for health states considered worse-than-dead** during the UO mode (77%) versus the RA mode (96%). No significant difference was observed in between the UO and F2F surveys.

- The results of the test-retest of the cTTO data showed no difference in the stability of preferences between MoA. For all modes test-retest correlation was categorised as "good".

Face-validity findings

- **Figure 2** shows an inverse relationship between utility value and sum score for all modes. Mean utility values elicited via F2F surveys were consistently lower than those elicited via the other modes, except for the worst health state.
- The largest, observable differences in mean utility values occurred 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 2. Mean health state utility elicited via cTTO plotted against severity



Conclusions

- The findings from this study suggest that data quality from DCE studies is not impacted by MoA.
  - **Eliciting preferences via DCE using either UO or interviewer-assisted MoA can be expected to provide similar, high quality data provided appropriately detailed instructions are provided to respondents**
- The quality of cTTO data differed substantially between UO and both F2F and RA modes, although the UO surveys were significantly quicker to complete than assisted modes.
- 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 the UO mode compared with either F2F or RA. Therefore, it may be more feasible to conduct cTTO valuations using EQ-VT via an interviewer-assisted mode than via UO surveys.
- The study findings suggest that the UO MoA provided insufficient guidance on how to complete the cTTO tasks, compared with modes in which an interviewer was present. Consideration could be given to providing additional guidance and interactive elements to improve engagement and explanation in UO surveys.
- 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.
  - **An interviewer-assisted MoA (either F2F or RA) is recommended for cTTO studies. A hybrid approach allowing respondents to select F2F or RA as preferred could be attractive economically if the cost of RA surveys is lower than F2F surveys**
- Comparing F2F and RA modes provided limited evidence to suggest either mode produces higher quality data than the other. Practical considerations should guide future researchers' choice of MoA between F2F and RA surveys.

References

1. Devlin NJ, Brooks R (2017) EQ-5D and the EuroQol Group: past, present, and future. Appl Health Econ Health Policy 15 (2):127-137
2. Oppe M, Devlin NJ, van Hout B, Krabbe PF, de Charro F (2014) A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health 17 (4):445-453
3. Oppe M, Rand-Hendriksen K, Shah K, Ramos-Goñi JM, Luo N (2016) EuroQol protocols for time trade-off valuation of health outcomes. PharmacoEconomics 34:993-1004
4. Janssen BMF, Oppe M, Versteegh MM, Stolk EA (2013) Introducing the composite time trade-off: a test of feasibility and face validity. Eur J Health Econ 14 (Suppl 1):S5-S13
5. Norman R, King MT, Clarke D, Viney R, Cronin P, Street D (2010) Does mode of administration matter? Comparison of online and face-to-face administration of a time trade-off task. Qual Life Res 19 (4):499-508
6. Shah KK, Lloyd A, Oppe M, Devlin NJ (2013) 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 14 (Suppl 1):S65-S73
7. Rowen D, Brazier J, Keetharuth A, Tsuchiya A, Mukuria C (2016) Comparison of modes of administration and alternative formats for eliciting societal preferences for burden of illness. Appl Health Econ Health Policy 14 (1):89-104.
8. Rowen D, Mukuria C, Bray N, Carlton C, Longworth L, Meads D, O'Neill C, Shah K, Yang Y (2022) Assessing the comparative feasibility, acceptability and equivalence of videoconference interviews and face-to-face interviews using the time trade-off technique. Soc Sci Med 309 (115227)
9. Determann D, Lambooij MS, Steyerberg EW, de Bekker-Grob EW, de Wit GA (2017) Impact of survey administration mode on the results of a health-related discrete choice experiment: online and paper comparison. Value Health 20 (7):953-960.
10. Jiang R, Shaw J, Mühlbacher A, Lee TA, Walton S, Kohlmann T, Norman R, Pickard AS (2021) Comparison of online and face-to-face valuation of the EQ-5D-5L using composite time trade-off. Qual Life Res 30 (5):1433-1444.
11. Mulhern B, Longworth L, Brazier J, Rowen D, Bansback N, Devin N, Tsuchiya A (2013) Binary choice health state valuation and mode of administration: head-to-head comparison of online and CAPI. Value Health 16 (1):104-113
12. Watson V, Porteous T, Bolt T, Ryan M (2019) Mode and frame matter: assessing the impact of survey mode and sample frame in choice experiments. Med Decis Making 39 (7):827-841.
13. Stolk E, Ludwig K, Rand K, van Hout B, Ramos-Goñi JM (2019) Overview, update, and lessons learned from the international EQ-5D-5L valuation work: version 2 of the EQ-5D-5L valuation protocol. Value Health 22 (1):23-30
14. Yang Z, Luo N, Oppe M, Bonsel G, Busschbach J, Stolk E (2019) Toward a smaller design for ED-5D-5L valuation studies. Value Health 22 (11):1295-1302
15. Oppe M, van Hout B (2017) The "power" of eliciting EQ-5D-5L values: the experimental design of the EQ-VT. EuroQol Research Foundation, Rotterdam, The Netherlands

Acknowledgments

- This research was funded by Bristol Myers Squibb
- This study would not have been possible without the respondents who were willing to take part in research surveys and interviews

Presented at ISPOR Europe 2022; November 6-9, 2022; Vienna, Austria, and virtual

Email: sarah.hill@phmr.com