PharmacoEconomics (2023) 41:329–338
https://doi.org/10.1007/s40273-022-01208-9

ORIGINAL RESEARCH ARTICLE

The Egyptian EQ-5D-5L Extensive Pilot Study: Lessons Learned

Sahar Al Shabasy[1] · Bram Roudijk[2] · Maggie Abbassi[1] · Aureliano Finch[2] · Elly Stolk[2] · Samar Farid[1]

Accepted: 6 October 2022 / Published online: 25 November 2022

Abstract

Objectives To investigate the effect of an extensive pilot phase in improving protocol compliance, face validity, reduction of interviewer effect and prediction errors in the composite time trade-off (cTTO) data elicited as part of the EQ-5D-5L valuation study in Egypt.

Methods This study used the cTTO data and quality control (QC) reports from the Egyptian EQ-5D-5L valuation study. Three-level mixed models were estimated to test whether interviewer effects were reduced during the pilot phase and subsequent rounds of collected cTTO data. Ordinary least square (OLS) regression analysis was conducted for each interviewer separately to test whether the mean absolute error (MAE) improved as interviewers completed more interviews. Moreover, improvement in protocol compliance, face validity and reduction of prediction errors in the cTTO data were tested.

Results 1180 interviews were conducted by nine interviewers and included in the final analysis, of which 206 interviews were pilot and 974 interviews were actual. There was substantial improvement in the face validity and reduction of prediction errors in the cTTO data where the MAE of the actual data was 0.37, which is much lower than that of the pilot data, which was 0.44. However, there was an initial high level of protocol compliance in terms of the four indicators of the QC tool and the variance attributed to the interviewers was small throughout the whole study.

Conclusions This study clarified the benefits of the pilot phase and the strict implementation of the QC tool in improving the face validity and the prediction accuracy of the cTTO data. However, a more extensive pilot phase may be more beneficial in EQ-5D-5L valuation studies that have issues initially with protocol compliance and interviewer effects.

1 Introduction

The EQ-5D is the most used multidimensional instrument for measuring health-related quality of life and quality-adjusted life years [1]. A partial explanation for the popularity of the EQ-5D is that many EQ-5D value sets are available that were constructed at the national level, reflecting the belief that preferences for health can differ across populations. The EuroQol group developed a standardized valuation protocol for the EQ-5D-5L valuation studies that implements two valuation techniques: the composite time trade-off (cTTO) and discrete choice experiment (DCE). Additionally, interviewer training materials are standardized and officially translated in an attempt to harmonize the methodology and training of interviewers in valuation studies across different countries to maximize comparability of the resulting value sets [2, 3]. Development of a country-specific value set using these valuation techniques is nevertheless challenging as it requires trained interviewers for guidance of participants through the whole interview process [4, 5]. Interviewer behavior might also add unwanted variability to the data.

The results of the first wave of valuation studies for the EQ-5D-5L raised the importance of data quality, especially in the cTTO part of the data collection. Multiple issues were observed including few worse than dead responses, low values for mild states, clustering of values and high frequency of inconsistent responses [6–9]. When The EuroQol realized that these issues were interviewer-driven, measures were taken to promote the performance of the interviewers [10, 11]. Refinements of the valuation protocol included the introduction of the quality control (QC) tool, feedback module and three practice states to improve the reliability and validity of the data and promote interviewer performance [10–12].

In the cyclic QC process, Ramos-Goni et al. defined minimum requirements for achievement of protocol compliance

✉ Samar Farid
samar.farid@pharma.cu.edu.eg

[1] Department of Clinical Pharmacy, Faculty of Pharmacy, Cairo University, Kasr El-Aini St., Cairo 11562, Egypt

[2] EuroQol Research Foundation, Rotterdam, The Netherlands

Key Points for Decision Makers

A pilot phase may increase the cost of valuation studies and is time consuming. It is currently unknown whether an extensive pilot phase has a meaningful impact on the performance of the interviewers and whether it may help minimize interviewer effects in EQ-5D-5L valuation studies.

This study highlighted the benefits of an extensive pilot phase on data quality and interviewer performance in terms of improvement in the face validity and reduction of prediction errors in the cTTO data during the whole data collection process and especially during the pilot phase.

A pilot phase may have substantial benefits for data collection of EQ-VT studies where it can help reveal issues and exclude poorly performing interviewers and might prove more beneficial in EQ-5D-5L valuation studies where protocol compliance issues and interviewer effects exist.

as baseline for the initial assessment for each interviewer to complete or stop data collection. The cyclic nature of the process allowed the study teams to reflect on interviewer's performance and gave them continuous feedback to improve their skills and minimize interviewer effects during the entire data collection period [10]. However, there are other factors such as sociodemographic characteristics of the participants and their preferences that might contribute to the apparent existence of interviewer effects [13, 14]. Since a pilot phase is not usually used in the EQ-5D-5L valuation studies, it is not clear whether an extensive pilot phase is required for the EQ-5D-5L valuation studies to improve data quality and standardize interviewers' performance.

The aim of this study is to investigate how interviewer performance evolved during the EQ-5D-5L valuation study in Egypt and to investigate the effect of the extensive pilot phase in improving protocol compliance, face validity, and reduction of interviewers' effect and prediction errors in the cTTO data. Identifying all these aspects can provide a guide for designing future valuation studies and training materials and help in improving the performance of interviewers and the quality of the collected data.

2 Methods

2.1 Data Source

This study used cTTO data and QC reports of the Egyptian EQ-5D-5L valuation study [15]. A total of 1,303 interviews were conducted in the period between July 2019 and March 2020 by 12 interviewers and two principal investigators (PIs). Ten interviews were test interviews done by the PIs. Once interviewers were recruited and trained, they did pilot interviews until the study team decided that they had acquired the necessary expertise to obtain good quality interviews based on the QC tool. Three interviewers were excluded due to interviewer effects seen in the data (113 interviews). The final analysis of this study included 206 pilot interviews and the 974 actual interviews that were included in calculating the Egyptian tariff [15]. Members of the general public were recruited from different Egyptian governorates using multi-stratified quota sampling to select a representative sample in terms of age, sex and geographical distribution. Each participant was interviewed face to face by a trained interviewer using the Egyptian translated version of the EQ-VT-2.1 protocol [2]. Interviews took place at the interviewers' office or the participants' home, workplace or other public places according to the participants' preferences. The interviewer training was performed in four stages: interviewing the interviewers by the PIs, initial training followed by conducting pilot interviews then retraining [16].

2.2 Quality Control (QC)

The QC reports are composed of two main aspects, namely protocol compliance and interviewer effects, in addition to other meta data such as the number of iteration steps and the time spent on the better than dead (BTD) and worse than dead (WTD) section of the cTTO task [10]. Protocol compliance is assessed based on four criteria such as the time spent on the WC example and actual cTTO tasks should not be less than 3 min and 5 min, respectively, the presence of clear inconsistency in the cTTO rating or if the interviewer did not use the lead time in the WC example. The interview was flagged if the interviewer was not compliant with any of the above-mentioned criteria. A conservative threshold of four flagged interviews out of ten was established as the limit to stop and retrain the interviewer, after a further ten interviews for the same interviewer, if again four or more interviews were flagged, the interviewer should be excluded from data collection [10]. Interviewer effects were assessed for any unusual clustering or distribution by comparing the cTTO value distribution for each interviewer to the overall distribution of values for all interviewers. The QC reports were discussed through periodical online meetings: weekly during the pilot phase (every five interviews per interviewer) and every 2 weeks during actual data collection (every ten interviews per interviewer) between the Egyptian team and the EQ-VT support team, and the feedback received was discussed with all interviewers. All 12 interviewers were

compliant with the minimum requirements of the protocol. However, three interviewers, along with the interviews they had conducted, were excluded from data collection process and data analysis due to the presence of strong clustering and inconsistent distributions for the cTTO data despite retraining and close monitoring, which could indicate poor engagement in the valuation tasks and interviewer' effects.

2.3 Data Analysis

Analyses were conducted using IBM SPSS Statistics for Windows, Version 22.0 (Armonk, NY, USA: IBM Corp) for sample demographic and QC indicators, STATA software version 14 was used to test for the protocol compliance, interviewer effects, clustering and predictive accuracy.

2.3.1 Sample Demographic Characteristics and QC Tool Indicators

Descriptive statistics were presented for sample socio-demographics and the QC tool indicators; we used percentages to present discrete variables, mean and standard deviation for continuous variables.

2.3.2 Protocol Compliance, Interviewer Effects and Clustering

Data were divided into batches of ten interviews by interviewer. We examined the rate at which interviews were flagged between the pilot phase and the actual data collection phase and calculated the rate of flagged interviews by interviewer to compare the effect of the pilot phase on improving protocol compliance, and to investigate whether the rate of flagged interviews decreased beyond the pilot phase or stopped decreasing within the pilot phase. This allowed the determination of whether there was a decreasing trend in flagged interviews along the study.

To test whether interviewer effects were reduced during the pilot phase and subsequent rounds of collected cTTO data, three-level mixed models were estimated where the variance in values was partitioned into variance attributed to responses, variance attributed to respondents, and variance attributed to interviewers by using responses nested in respondents, nested in interviewers on each of the subsamples of ten interviews per interviewer per batch. Intraclass correlation (ICC) coefficients were calculated to investigate whether there was a decreasing trend in the share of variance attributed to interviewers over the collected rounds of data.

Reduction of clustering on the easily obtained values such as (-1, -0.5, 0, 0.5 and 1) were compared and taken as an initial indication of quality improvement. Scatter plots were used to investigate whether clustering decreased over rounds of the collected data.

2.3.3 Predictive Accuracy

To test whether the pilot phase had a significant effect on the aggregate predictive accuracy of the models employed in the value set calculation, two samples were compared, the sample used for the value set calculation ($n = 974$) and a sample of equal size including the pilot data ($n = 206$) and the first 768 actual interviews. The omission of actual interviews in the second sample was balanced by interviewer, where the numbers of actual interviews excluded for each interviewer were equal to their pilot interviews. First, we applied the Egyptian value set to all health states valued in the pilot and actual data [15]. For each of the two samples, the mean absolute error (MAE) was computed by comparing the mean of the difference between the values assigned by respondents and the index values. As a comparison, we did a random draw of similar size of two other samples out of all data collected, pilot and actual data, and their performance was compared with that data. The random draws were repeated 10,000 times, to ensure robustness of the sample selection.

To determine whether the pilot data caused better predictive accuracy at the interviewer level after doing more interviews, the Egyptian value set was applied to the valuation data. Then, we calculated the MAE within each interview (ten responses per interview) by taking the mean of the difference between the index values and the values provided by the respondents. Subsequently, using scatter plots, decreasing trends in the MAE over time were visualized by plotting the MAEs within each interviewer over the sequence of interviewing.

Ordinary least square (OLS) regression analysis, with the respondent-level MAE as the dependent variable and the rank order in which the interviews were conducted by the interviewer (Time) as the independent variable, were conducted for each interviewer separately (Eq. 1). This allowed us to test whether the MAE improves when interviewers complete more interviews, in other words, whether the outcomes of a cTTO interview become more similar to the results of the final value set. In addition, we explored models that included a dummy variable (Pilot) that indicated whether data was pilot data (Pilot = 1) or non-pilot data (Pilot = 0) (Eq. 2), and also the interaction between whether the data are pilot data and the sequence of interviews (Time*Pilot) (Eq. 3). For each of these variables, p-values were calculated to test the significance of their relationship with the respondent-level MAE. Significant parameter estimates for the dummy variable showed that the MAE was larger or smaller in the pilot, compared to the actual data, and the interaction term showed whether the improvement in predictive error was larger in the pilot phase:

$$\mathrm{MAE}_i = \beta_0 + \beta_1 \mathrm{Time} + \varepsilon_i, \quad (1)$$

$$\mathrm{MAE}_i = \beta_0 + \beta_1 \mathrm{Time} + \beta_2 \mathrm{Pilot} + \varepsilon_i, \quad (2)$$

$$\mathrm{MAE}_i = \beta_0 + \beta_1 \mathrm{Time} + \beta_2 \mathrm{Pilot} + \beta_3 \mathrm{Pilot} * \mathrm{Time} + \varepsilon_i. \quad (3)$$

In Eqs. (1), (2) and (3), MAE_i represents the mean absolute error for the interview conducted with respondent i. β_0 represents the regression intercept, while β_1Time represents the effect of interview sequence. β_2Pilot and β_3Pilot $*$ Time represent the effect of the pilot phase and the interaction between the pilot phase and interview sequence, respectively. ε_i is the residual variance.

3 Results

3.1 Sample Demographic Characteristics

In this study 1180 interviews were included in the final analysis; these interviews were conducted by nine interviewers who completed data collection, of which 206 interviews were pilot and 974 interviews were actual.

Table 1 gives an overview of the study sample characteristics. The majority of the participants in the pilot phase were highly educated, employed and lived in urban areas in Cairo.

3.2 QC Tool Indicators

Table 2 compares the QC tool indicators for the pilot and actual data collection phases showing the improvement in the actual data collection phase.

3.3 Protocol Compliance, Interviewers' Effects and Clustering

Data were divided into 14 batches clustered by interviewer. The first three batches represented the pilot phase ($n = 206$) and the subsequent batches (4–14) represented the actual data collection phase ($n = 974$). The average number of interviews per interviewer in the pilot phase was 23 (range 10–40), and the average number of interviews in the actual data collection was 108 (range 78–169). Each batch consisted of ten interviews per interviewer, except batches 3 and 14. Table 2 shows the exact number of pilot and actual interviews for each interviewer.

There was no effect of the pilot phase on protocol compliance in terms of the four indicators of the QC tool, where the percentages of flagged interviews did not exceed 3.3% per batch in the pilot phase nor in the actual data collection phase. There was no improvement in interviewer effects beyond the pilot phase and it did not decrease substantially over time. However, the share of variance attributed to interviewers over the collected rounds of data, as demonstrated by ICC, did not exceed 6.7% through the whole study.

There was improvement in the face validity of the data where less clustering over time was observed in the easily attained responses (Fig. 1a). In addition, in the pilot phase the range of the mean number of unique values per respondent was 5.7–6.3, which increased to 6.9–8.1 in the actual data collection phase (Fig. 1b). Moreover, the percentages of respondents with fewer than five unique values decreased through the data collection process where the range was 16.7–25.6 in the pilot and 3.3–12.6 in the actual data collection.

The percentage of respondents only using integers in trading the life years decreased through the data collection process where the range decreased from 36.8–46.7 to 13.6–40 for the pilot and actual data respectively (Fig. 2a).

3.4 Predictive Accuracy

The predictive accuracy increased over batches and beyond the pilot phase where the range of MAE between the pilot and actual data per batch were 0.42–0.46 and 0.32–0.40, respectively (Fig. 2b). The MAE averaged across batches

Table 1 Background characteristics of the Egyptian participants

Characteristics	Actual data ($n = 974$)	Pilot data ($n = 206$)
Sex		
Male	510 (52.4)	100 (48.5)
Age (years)	36.9 ± 12.7 (18–72)	32.3 ± 12.5 (18–75)
18–34	450 (46.2)	148 (71.8)
35–54	420 (43.1)	38 (18.4)
≥ 55	104 (10.7)	20 (9.7)
Geographical region[a]		
Greater Cairo	256 (26.3)	162 (78.6)
Other regions	716 (73.7)	44 (21.4)
Residence[a]		
Urban	658 (67.7)	178 (86.4)
Education level[a]		
Illiterate	109 (11.2)	2 (1)
Below intermediate[b]	290 (29.8)	13 (6.3)
Intermediate[c]	398 (40.9)	64 (31.1)
University degree and above	175 (18)	127 (61.7)
Employment status[a]		
Employed	728 (74.9)	142 (68.9)

Data are presented as n (%) or mean ± standard deviation (range)

[a]Sample size was $n = 972$ for the actual data

[b]Below intermediate: below high school level

[c]Intermediate: high school level or 2 years institute

Table 2 Quality control (QC) tool indicators

QC tool indicators	Actual sample ($n = 974$)	Pilot sample ($n = 206$)
Flagged interviews *n* (%)	11 (1.1)	5 (2.4)
% of flagged interviews per interviewer (range)	0–4	0–11
Wheelchair example (mean ± SD)		
Total time (s)	214.6 ± 139.8	233.3 ± 138.9
Time on BTD element (s)	174.3 ± 144.6	186.4 ± 124.7
Time on WTD element (s)	40.3 ± 59.6	46.9 ± 79.8
Total moves	9.5 ± 4.1	10.5 ±6.9
Moves on BTD element	6.5 ± 4.3	7.2 ± 6.4
Moves on WTD element	3.0 ± 3.9	3.3 ± 4.5
Clustering (%)		
− 1	13.3	20.6
− 0.5	4	4
0	1.5	4.4
0.5	5.2	9.7
1	12.3	12.8
Time spent in feedback module, s (mean ± SD)	167.8 ± 624.2	193.3 ± 95.2
Total interview time, min (mean ± SD)	41.1 ±16.2	47.5 ±13.6

BTD better than dead, *min* minutes, *QC* quality control, *SD* standard deviation, *s* seconds, *WTD* worse than dead

of the actual data was 0.37, which is lower than that of the pilot data 0.44.and, the MAE for the first 974 interviews by including all the pilot interviews and the first 768 actual interviews was 0.39. Drawing 1000 respondents randomly from the whole dataset (pilot+actual data) (Fig. 3) lead to MAEs that were higher than the actual data, but lower than the pilot data. Figure 4 showed the MAEs within each interviewer over the sequence of interviewing by interviewer per respondent. It is clear from Fig. 4 that the noise in the data decreases in later rounds of interviews.

In Table 3 model A shows the OLS regression analyses for MAE over interview sequence (Time) by interviewer, where there was a significant effect for time for six out of nine interviewers, that proved that the MAE for most interviewers decreased once they did more interviews (sequence effect).

In model B adding the Pilot variable increased the explained variance (R^2) when compared to the (R^2) in model A, but the effect for the interaction variable (Time*Pilot) was not significant for most interviewers, as demonstrated by (pTime*Pilot) (model C). This is a

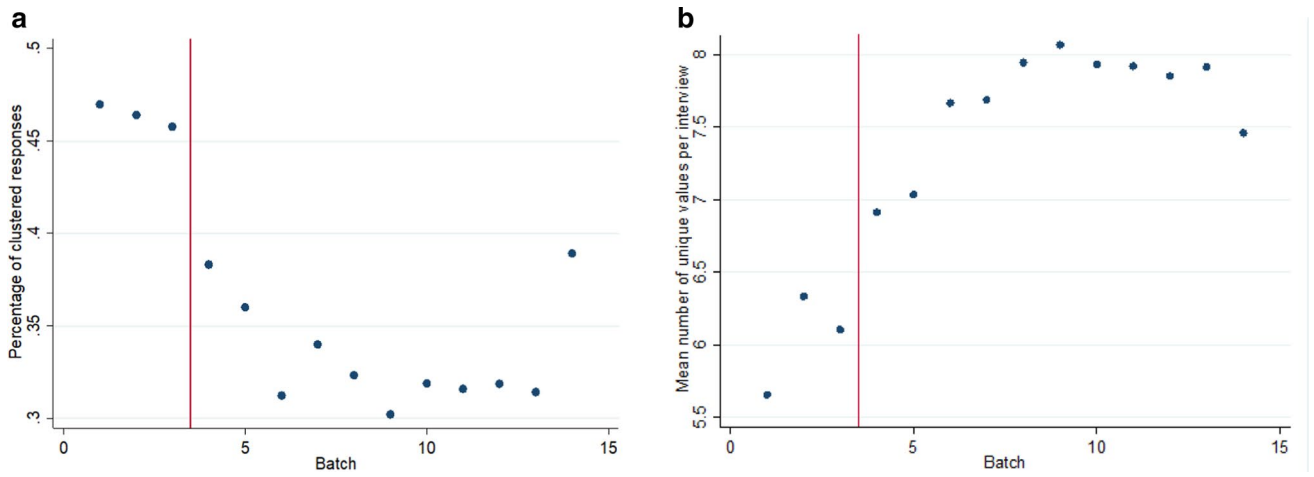

Fig. 1 Percentage of clustered responses (**a**) and mean number of unique values per interview (**b**) per batch, pilot data (left of the red line, actual data (right of the red line)

signal that the MAE was generally lower in the actual data compared to the pilot data as demonstrated by the signs of coefficients of the Pilot variable. This was compatible with the notion that the final value set model was estimated on the actual data, and the estimates for MAE for the pilot data were therefore an out-of-sample prediction, which was expected to have more error than a within-sample prediction. In model B, five out of nine interviewers still had a significant effect for time, which proved that regardless of whether the data were pilot data or not, the MAE was decreasing as the interviewers did more interviews. This showed that interviews completed provide responses that are more similar to the final value set model compared to the responses earlier in the study, which suggests that the precision of the interviews may have improved.

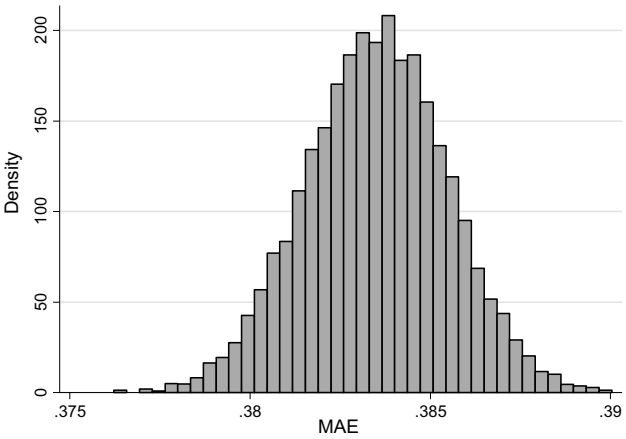

Fig. 3 Mean absolute errors (MAEs) for 10,000 random draws of 1000 respondents completing all ten composite time trade-off (cTTO) tasks

4 Discussion

4.1 Main Findings

To our knowledge this is the first study that highlighted the benefits of an extensive pilot phase on data quality and interviewer performance. We examined the improvement in protocol compliance, face validity and interviewer effects, in addition to the reduction of prediction errors in the cTTO data. Our main findings show that the face validity of the data seems to improve; that is, the number of unique values per respondent, as well as the use of non-integer numbers seems to increase, while clustering of values seems to decrease in the interviews included in the actual data collection versus the pilot phase. Furthermore, we have shown that the values collected in the pilot study are different from those collected in the actual data collection, as shown by the higher MAEs. The MAE seems to decrease for interviews conducted later in the data collection, both within the pilot phase as well as in the interviews completed as part of the actual data collection.

4.2 Interpretation

The face validity and prediction error data show a similar pattern; during the pilot phase there is a substantial improvement in the key characteristics examined in the current study, due to the feedback shared with the interviewers regarding their performance, where a written debriefing was sent to each interviewer that included formative evaluation of their performance and the main issues to be considered during

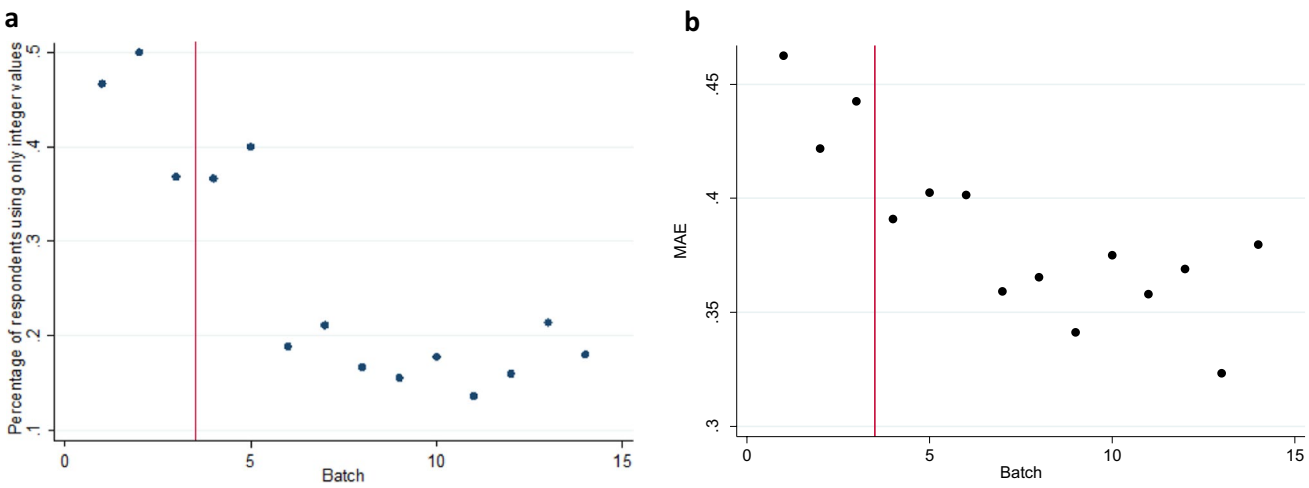

Fig. 2 Percentage of respondents using only integer values (**a**), mean absolute error (MAE) (**b**) per batch, pilot data (left of the red line, actual data (right of the red line)

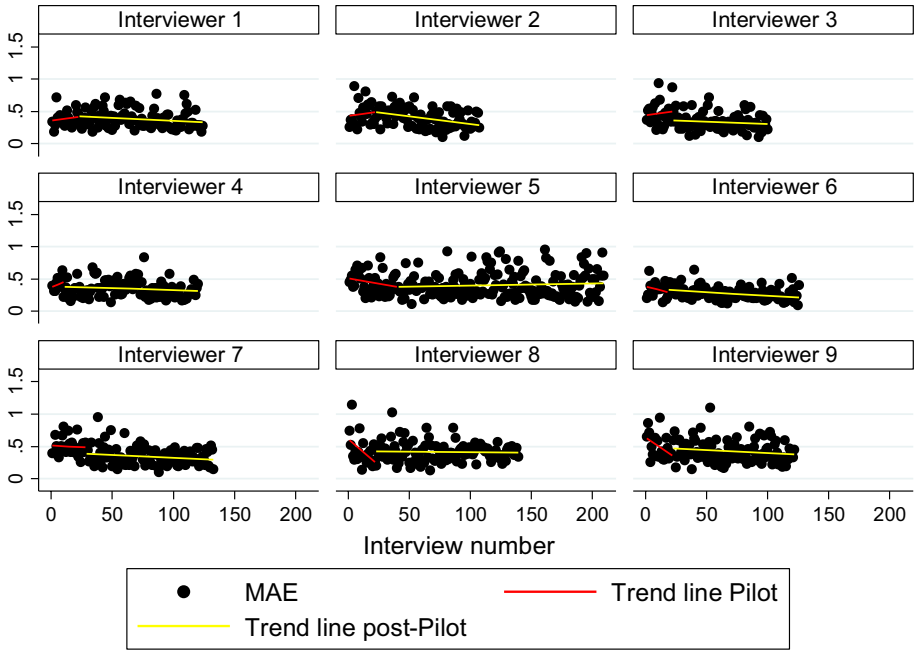

Fig. 4 Mean absolute error (MAE) per respondent per interviewer over the sequence of interviews

Table 3 Regression coefficients for mean absolute error (MAE) (model A) over interview sequence (Time) by interviewer, (model B) type of data (pilot or actual data) and (model C) interaction between interview sequence (time) and type of data

Interviewer	1	2	3	4	5	6	7	8	9
Number of interviews									
Pilot	22	22	22	10	40	18	28	22	22
Actual	102	85	78	110	169	108	105	117	100
Total	124	107	100	120	209	126	133	139	122
Model A: Interview sequence (Time)									
Time	− 0.001	− 0.002	− 0.002	− 0.001	0.000	− 0.001	− 0.002	− 0.000	− 0.001
Intercept	0.4167	0.504	0.457	0.405	0.4097	0.348	0.486	0.433	0.505
pTime	0.089	**0.000**	**0.000**	**0.012**	0.928	**0.000**	**0.000**	0.449	**0.013**
R^2	0.024	0.135	0.123	0.052	0.000	0.144	0.176	0.004	0.050
Model B: Pilot (type of data, pilot or actual data)									
Time	− 0.001	− 0.002	− 0.001	− 0.001	0.000	− 0.001	− 0.001	− 0.000	− 0.001
Pilot[a]	− 0.043	− 0.052	0.102	0.022	0.073	− 0.000	0.097	− 0.008	0.003
Intercept	0.443	0.541	0.381	0.398	0.362	0.348	0.417	0.437	0.503
pTime	0.050	**0.000**	0.272	**0.049**	0.246	**0.000**	**0.035**	0.490	**0.07**
pPilot	0.297	0.276	**0.036**	0.633	0.109	0.994	**0.019**	0.867	0.957
R^2	0.032	0.1446	0.1617	0.054	0.012	0.144	0.211	0.004	0.050
Model C: Interaction between interview sequence (time) and pilot (type of data pilot or actual data)									
Time	− 0.001	− 0.002	− 0.001	− 0.001	0.000	− 0.001	− 0.001	− 0.000	− 0.001
Pilot[a]	− 0.076	− 0.127	0.048	− 0.035	0.148	0.035	0.098	0.176	0.129
Time*Pilot	0.003	0.006	0.004	0.010	− 0.003	− 0.004	− 0.000	− 0.015	− 0.010
Intercept	0.445	0.547	0.387	0.398	0.356	0.347	0.417	0.429	0.495
pTime	**0.045**	**0.000**	0.223	**0.047**	0.192	**0.001**	**0.037**	0.643	0.102
pPilot	0.260	0.100	0.535	0.697	**0.042**	0.490	0.138	**0.022**	0.138
pTime*Pilot	0.533	0.216	0.362	0.453	0.185	0.392	0.996	**0.003**	0.068
R^2	0.035	0.157	0.169	0.058	0.021	0.149	0.211	0.068	0.077

Bold values are significant at a p value < 0.05

[a]Pilot, coded as 1= pilot data, 0 = actual data

the next set of interviews. In addition, the interviewers were advised to standardize the outline of the interview during the cTTO task, to ensure precision of responses. This included informing the respondents that they were presented with different health states with different severity levels and to show them the full range of the TTO scale with the 6-month increments or decrements during the example questions. In addition, they were instructed to ask the participants for the rationale of their answers if their answers were illogical. Furthermore, the MAE also seems to improve within the interviewer completing more interviews. These two outcomes combined suggest that there is a learning effect (sequence effect) for the interviewers, leading to better data quality after the pilot phase. After completing the pilot phase, there were still some improvements, but not as large as the improvements made during the pilot phase. This may suggest a role for the implementation of pilot phases in future EQ-VT studies.

The MAE data show that there is a substantial difference between the MAE of the data used for the Egyptian value set (0.37) and the pilot data (0.44). Although the MAE for the pilot data is based on an out-of-sample prediction, one may still expect the difference in MAE to be very small if one expects no effect of a pilot phase on predictive accuracy. This along with the observation that MAEs on average decrease by individual interviewer over their interview sequence, strengthens the observation that a pilot phase has a positive effect on the predictive accuracy of the collected data.

The Egyptian valuation study showed high levels of protocol compliance in terms of the four indicators of the QC tool during the initial waves of data collection in the pilot phase where the percentage of flagged interviews per interviewer did not exceed 11% per interviewer in the pilot phase, and 4% in the actual data collection phase. In other studies, this is typically higher, for example the Peruvian EQ-5D-5L valuation study reported 0–19% of interviews flagged per interviewer [17]. This might be attributed to using the QC tool elements as part of the training of interviewers for the Egyptian valuation study. It seems that the protocol compliance was initially already high, which means the effect of a pilot phase on protocol compliance may have been limited in the current study. However, studies that initially report lower rates of protocol compliance may possibly still improve protocol compliance rates during a pilot phase before actual data collection is started.

In EQ-5D-5L valuation studies, interviewers have a major role in motivating respondents to engage in the valuation tasks and to express their values accurately, in addition to dealing with certain participant behaviour or characteristics. In this study, the interviewer training was extensive and performed in four stages to minimize inter- and intra-interviewer effects and to improve performance; this process has been detailed in a previous publication [16]. There was no improvement in interviewer effects beyond the pilot phase and it did not decrease substantially over time. However, the share of variance attributed to interviewers did not exceed 6.7% through the whole study. This might be attributed to the difference in interviewers' personalities and style in addition to the variation in the characteristics of the participants, time and place of the interview (regional difference of values), which might have an impact on how participants completed the valuation interview [13, 14, 17]. Other studies report interviewer effects as well, but did not quantify them as reported in the current study, making it difficult to make comparisons [17–19].

Overall, it seems like a pilot phase may have substantial benefits for data collection of EQ-VT studies. From our data, we show that there is a likely learning effect, where the quality of the collected data increased with the number of interviews completed by an interviewer, the more interviewing experience the higher the level of prediction accuracy and lower level of logical inconsistency. The lower number of inconsistent responses reported when interviewers are more experienced was also found in a previous study by Yang et al. [20]. The lessons learned from the extensive pilot phase in the Egyptian valuation study and the strict implementation of quality control allowed us to provide the interviewers with better feedback, which improved their performance. Although all these requirements increased study costs and led to removal of data, implementing an extensive pilot phase seems to be very effective at revealing data quality issues, and improving the quality of the sample used for estimating the value set.

4.3 Strengths and Limitations

This is not the first EQ-VT study in which a pilot phase was implemented before the final data collection phase commenced. However, it is the first in its current structure, where in our study, each interviewer completed an average of 23 pilot interviews before they commenced actual data collection. In Peru and France the interviewers conducted only five to ten pilot interviews [17, 21]. This is substantially less than in our study, and the size of the pilot sample allowed us to assess the effects of a pilot phase in more detail than possible in previous studies, which is a strength of this study.

One of the limitations of this study is that there were some differences in the sample background characteristics between the pilot phase and the actual data collection phase, where most of the participants in the pilot phase were highly educated, employed, and lived in urban areas in Cairo; it is usually preferred for the pilot study to take place in a central location to reduce cost, achieve a consistent sample frame for all interviewers, and facilitate PI-interviewer

interactions. However, it is not clear how this has affected the results.

4.4 Implications

For EQ-5D-5L valuation studies, achieving the minimum quality control requirements is not enough to guarantee good data quality. As shown in the current study, implementing an extensive pilot may substantially improve the face validity and predictive accuracy of the data collected in the actual data collection phase, which may guarantee the highest standards of data quality for generating value sets. Moreover, interviewer effects should be more carefully addressed particularly in the QC process with the development of more exploratory research to control interviewer effects in future EQ-5D-5L valuation studies.

5 Conclusion

This study clarified the benefits of the pilot phase and the strict implementation of the QC tool in improving the face validity and the prediction accuracy of the cTTO data. However, a more extensive pilot phase may be more beneficial in EQ-5D-5L valuation studies that initially have more issues with protocol compliance and interviewer effects.

Acknowledgements We would like to thank the EuroQol support team for their guidance and support in the study preparation, data collection and quality control process. Special thanks to the interviewers for their outstanding work and assistance.

Declarations

Funding This project received financial support from the EuroQol Research Foundation, The Netherlands (EQ-Project 335-RA). The funding agreement ensured the authors' independence in designing, writing and publishing the study results.

Conflict of interest Aureliano Paolo Finch, Bram Roudijk and Elly stock are members of the EuroQol research foundation (the copyright holder of the EQ-5D-5L). Sahar A. Al Shabasy, Maggie M. Abbassi, and Samar F. Farid have no conflicts of interest directly relevant to the content of this article.

Availability of data and material The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Ethics approval The study received ethics approval from the Ethics Committee of the Faculty of Pharmacy Cairo University and was conducted in accordance with the Declaration of Helsinki.

Consent to participate Written informed consent was obtained from all participants included in the study. Participants were informed about their freedom of refusal. Anonymity and confidentiality were maintained throughout the research process.

Consent for publication The authors, jointly give the publisher permission to publish this work.

Code availability Not applicable.

Author contributions SAS participated in the study preparation and data collection, created data QC reports, interpreted results and prepared the draft manuscript. BR participated in the statistical analysis, interpretation of results and review of the final manuscript. MA and SF participated in the study preparation, proof-reading of the translated version, follow-up of the data collection, the QC process, interpreting results and reviewing the final manuscript. AF participated in follow-up of data collection, the QC process and review of the final manuscript. ES participated in the study preparation and reviewed the final manuscript.

References

1. Wisløff T, Hagen G, Hamidi V, Movik E, Klemp M, Olsen JA. Estimating QALY gains in applied studies: a review of cost-utility analyses published in 2010. Pharmacoeconomics. 2014;32(4):367–75.
2. 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.
3. Devlin NJ, Krabbe PF. The development of new research methods for the valuation of EQ-5D-5L. Eur J Health Econ. 2013;14(Suppl. 1):1–3.
4. Lugnér AK, Krabbe PF. An overview of the time trade-off method: concept, foundation, and the evaluation of distorting factors in putting a value on health. Expert Rev Pharmacoecon Outcomes Res. 2020;20(4):331–42.
5. Edelaar-Peeters Y, Stiggelbout AM, Van Den Hout WB. Qualitative and quantitative analysis of interviewer help answering the time tradeoff. Med Decis Mak. 2014;34(5):655–65.
6. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: an EQ-5D-5L value set for England. Health Econ. 2018;27(1):7–22.
7. Versteegh MM, Vermeulen KM, Evers SM, De Wit GA, Prenger R, Stolk EA. Dutch tariff for the five-level version of EQ-5D. Value Health. 2016;19(4):343–52.
8. Ramos-Goñi JM, Pinto-Prades JL, Oppe M, Cabasés JM, Serrano-Aguilar P, Rivero-Arias O. Valuation and modeling of EQ-5D-5L health states using a hybrid approach. Med Care. 2017;55(7):e51–8.
9. Alava MH, Pudney S, Wailoo A. The EQ-5D-5L value set for England: findings of a quality assurance program. Value Health. 2020;23(5):642–8.

10. 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.
11. 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.
12. Purba FD, Hunfeld JA, Iskandarsyah A, Fitriana TS, Sadarjoen SS, Passchier J, et al. Employing quality control and feedback to the EQ-5D-5L valuation protocol to improve the quality of data collection. Qual Life Res. 2017;26(5):1197–208.
13. Al Sayah F, Bansback N, Bryan S, Ohinmaa A, Poissant L, Pullenayegum E, et al. Determinants of time trade-off valuations for EQ-5D-5L health states: data from the Canadian EQ-5D-5L valuation study. Qual Life Res. 2016;25(7):1679–85.
14. Al Shabasy S, Al Sayah F, Abbassi M, Farid S. Determinants of health preferences using data from the Egyptian EQ-5D-5L Valuation Study. Patient. 2022;15(5):589–98.
15. Al Shabasy S, Abbassi M, Finch A, Roudijk B, Baines D, Farid S. The EQ-5D-5L valuation study in Egypt. Pharmacoeconomics. 2022;40(4):433–47.
16. Al Shabasy S, Abbassi M, Farid S. EQ-VT protocol: one-size-fits-all? Challenges and innovative adaptations used in Egypt: a cross-sectional study. BMJ Open. 2021;11(12): e051727.
17. Augustovski F, Belizán M, Gibbons L, Reyes N, Stolk E, Craig BM, et al. Peruvian valuation of the EQ-5D-5L: a direct comparison of time trade-off and discrete choice experiments. Value Health. 2020;23(7):880–8.
18. Pattanaphesaj J, Thavorncharoensap M, Ramos-Goñi JM, Tongsiri S, Ingsrisawang L, Teerawattananon Y. The EQ-5D-5L valuation study in Thailand. Expert Rev Pharmacoecon Outcomes Res. 2018;18(5):551–8.
19. Lin H-W, Li C-I, Lin F-J, Chang J-Y, Gau C-S, Luo N, et al. Valuation of the EQ-5D-5L in Taiwan. PLoS ONE. 2018;13(12): e0209344.
20. Yang Z, van Busschbach J, Timman R, Janssen M, Luo N. Logical inconsistencies in time trade-off valuation of EQ-5D-5L health states: whose fault is it? PLoS ONE. 2017;12(9): e0184883.
21. Andrade LF, Ludwig K, Goni JMR, Oppe M, de Pouvourville G. A French value set for the EQ-5D-5L. Pharmacoeconomics. 2020;38(4):413–25.