Behind the Scenes: A Qualitative Investigation of Interviewers' Performance in EQ-5D Valuation Studies

Carlotta Varriale[1] · Giovanni Andrulli[2] · Michela Meregaglia[3] · Fanni Rencz[4] · Aureliano Paolo Finch[5,6]

Abstract

Background The EuroQol Valuation Technology (EQ-VT) protocol is currently employed by the valuation studies of the EQ family of instruments worldwide. To date, all the evidence in support of the quality control (QC) originates from quantitative indicators.

Objective We aimed to explore interviewers' conversational patterns in EQ-VT interviews, beyond quantitative QC indicators, and to provide a preliminary exploration of how the interaction between interviewer and respondent impacts data quality.

Methods Two researchers transcribed and independently coded 24 video-recorded interviews from the Italian EQ-5D-5L valuation study, adopting the conversational analysis framework. The analysis identified positive and negative 'patterns' of conversational practice. These were categorized into themes and sub-themes and were used to score a random sample of 42 video-recorded interviews conducted at different time points by seven interviewers.

Results The conversational analysis identified 20 positive and 14 negative interview patterns, which were grouped into two main themes (i.e., task execution and communication skills). Positive items included appending questions that stimulated respondents' engagement, providing different explanations for an unclear aspect, supporting the participant with useful information for completing the tasks, and increasing the interview's coherence by confirming the respondent answers. Negative patterns included moving forward in the exercise without making sure that the respondent understood, trying to force an answer from the respondent, speaking too fast, and providing incomplete or incorrect explanations of the task. Most interviewers exhibited a moderate increase in positive patterns or a decrease in negative patterns over time. A certain degree of consistency between the quantitative QC results and the qualitative scoring deriving from the interviewer-respondent interaction was observed, with the best and worst performers of the qualitative scoring showing good and bad scores on key QC items, respectively.

Conclusions The identified positive and negative patterns may be useful to inform the training material of EQ-VT studies worldwide and complement the existing QC process.

✉ Michela Meregaglia
  michela.meregaglia@unibocconi.it

1 Economics Department, OECD, Paris, France

2 Lamb Weston Meijer, Breda, The Netherlands

3 CERGAS, SDA Bocconi School of Management, Milan, Italy

4 Department of Health Policy, Corvinus University of Budapest, Budapest, Hungary

5 EuroQol Office, EuroQol Research Foundation, Rotterdam, The Netherlands

6 Health Values Research and Consultancy, Amsterdam, The Netherlands

Key Points for Decision Makers

Quality control (QC) is a key component of the EuroQol Valuation Technology (EQ-VT) protocol, but thus far has included quantitative indicators only.

This study explores qualitative indicators of interviewers' performance in EQ-VT interviews by analysing interviewer-respondent interactions.

The conversational analysis identified 20 positive and 14 negative interview patterns grouped into 'task execution' and 'communication skills'.

△ Adis

1 Introduction

The EQ-5D-5L is one of the most widely used generic preference-based measures (GPBMs) of health [1, 2]. It comprises a descriptive system and a 20-cm vertical rating scale of self-reported health status i.e., EQ-VAS. The descriptive system describes health in terms of five dimensions and five levels, the combination of which results in 3125 ($=5^5$) health states. The value set assigns values to each of the health states described. Numerous studies have investigated the psychometric properties of the instrument, finding it is valid and responsive in multiple disease areas and conditions [3, 4].

Value sets (or tariffs) for the EQ family of instruments are obtained through valuation studies conducted among members of the general public of the country for which the value set is being generated. All studies for the EQ-5D-5L follow the EuroQol Valuation Technology (EQ-VT) protocol [5], which aims at promoting the standardization of valuation studies at an international level so that a similar valuation interview experience is offered to each respondent in every country and across countries.

Based on the protocol, respondents are shown two preference elicitation tasks, the composite time trade-off (cTTO) and the discrete choice experiment (DCE). The former represents the backbone of the value set, and respondents are familiarized with it using two wheelchair examples and three practice state questions. After completion of the familiarization session, 10 real cTTO questions are administered, followed by seven DCE tasks. Each interviewer receives an instruction manual and a standardized 2-day training session that illustrates the content and process of a valuation interview, the required standards of interaction between interviewers and respondents, and the key features of the quality control (QC) of EQ-VT studies.

The QC procedures are a key component of the EQ-VT protocol [6]. Following observations of high rates of inconsistent responses, clustering of values, low values for mild states, and a few worse-than-dead responses in the first wave of studies [7, 8], an in-depth analysis of valuation data was performed [6], which resulted in the identification of meta-indicators of poor data quality. The QC report broadly consists of a protocol compliance table, which is based on these indicators and shows, for example, the time spent in the wheelchair example, data distribution per interviewer and for all interviewers, and face validity of the aggregate data. The QC process is implemented by periodical monitoring of data collection rounds (usually in batches of 10 interviews per interviewer), which occurs on scheduled calls between two EQ-VT support members, and the Principal Investigators (PIs) of the valuation study. In these calls, the VT support members provide feedback to

the PIs on their interviewers' performance, which is then passed on to the interviewers in groups or individual calls.

As mentioned, the QC indicators were developed based on the analysis of aggregate cTTO data, and the implementation of QC resulted in clear improvements in those indicators [6, 9]. Recently, an influential group of researchers warned about the risks associated with the QC process [10]. Broadly, the argument relates to the fact that interviewers are subject to normal human behaviors and to the desire to conform/comply with indications. The organization of QC calls between the EQ-VT support team and PIs may lead interviewers to force respondents to 'confirm' their answers, as well as to form expectations around what is an expected response. Interviewer training and QC largely concern the soft skills of interviewers, such as verbal and non-verbal communication.

Nonetheless, these qualitative aspects of the valuation interviews have never been investigated, with all evidence in support of the QC originating from quantitative indicators. The practical impact of the QC process remains unknown.

The purposes of our study were to (1) explore the conversational patterns of interviewers' performance in EQ-VT interviews that are beyond what is measured during the QC process; and (2) provide a preliminary exploration of how the interaction between interviewer and respondent impacts data quality in EQ-VT interviews.

2 Methods

2.1 Data Collection and Study Population

This study was conducted as part of the larger Italian EQ-5D-5L valuation study [11, 12] approved by the Ethics Committee of Bocconi University in October 2020 (approval number: 2020-SA000136.4). In the Italian valuation study, 1183 interviews were collected between October 2020 and February 2021 by 11 interviewers via videoconferencing administration, following the EQ-VT protocol [13].

Data collection was structured in batches of 10 interviews per interviewer, after which a QC call was conducted to provide feedback to the interviewers. A round of 10 practice interviews preceded the first official batch. Further information on the Italian EQ-5D-5L valuation study is available in the study by Finch et al. [11]. In total, 152 interviews were video recorded, of which 8 were from the pilot phase, 63 from the first batch, 57 from the second batch, and 24 were from the remaining batches of data collection (Online Resource Table A.1).

2.2 Transcription and Coding of Interviews

Two authors (CV and GA) transcribed verbatim, in Italian, 24 video recordings from two interviewers starting from the wheelchair example until the fifth non-practice cTTO task. Six video recordings per interviewer were related to the first batch of data collection, and six were related to the second batch of data collection. The two interviewers were selected as they exhibited substantially different time stamps, number of moves, and data distribution based on the first QC assessment and based on the availability of recordings.

This selection was intended to maximize the likelihood of identifying positive and negative patterns of conversation. The 24 transcribed interviews were then coded using the Conversational Analysis framework [14, 15]. This approach has been previously employed in multiple disciplines, including health economics and health services research [16–18], and to assess the dyadic interaction between interviewers and respondents in standardized interviews [19].

2.3 Development of a Scoring System and Scoring of Interviews

The conversational analysis complemented by patterns previously identified in the literature [20, 21] lead to the development of a scoring system composed of Likert scales (1–3 or 1–5) and binary variables (yes/no items) based on the characteristics and the scope of each. Likert scales 1–5 were used for items showing relatively more frequencies than items expressed as Likert scales 1–3.

For the 1–5 frequency scales, we considered a score above 4 as a good result for a positive check and the same score (≥ 4) as a bad result for a negative check. For the 1–3 frequency scales, 1 was considered as a bad result for a positive check, 2 as a neutral result, and 3 as a good result, and the opposite for negative checks. Scorers promptly discussed all instances of disagreement among them and developed a set of rules to improve alignment when relevant, as reported in Online Resource Table A.5.

To calculate a final score per interview and interviewer, as given by Scorer 1 and 2, and assess its development over time, the Likert and yes/no (Boolean) items were standardized and a weighted average of positive and negative items was computed. To compute such averages we proceeded with the following steps. First, we separated positive and negative items to find a total score for positive items and a total score for negative items. Second, Likert scale values were divided by the maximum value of each scale, e.g. Likert scales 1–3 were divided by 3 and Likert scales 1–5 were divided by 5. In this way, the maximum value for each type of Likert scale became 1. Third, an average for each scale was computed (to obtain an average ranging from 0 to 1 for Likert 1–3, an average for Likert 1–5, etc., for both

positive and negative patterns). Fourth, a weighted average of the three scales was computed, weighted by the number of items for each scale (e.g., the positive checklist comprised 14 items in Likert 1–3, 5 items in Likert 1–5, and 1 Boolean item). This ensured that we obtained a weighted positive score and a negative score, where all the items have the same weight. In this way, we obtained an overall positive and overall negative score on a range from 0 to 1.

Overall, 42 video recordings were selected for the scoring phase from the 152 video-recorded interviews, excluding the 24 that had been used to develop the scoring system. Four interviewers (out of 11) were not included as there were limited video recordings for them. For the remaining seven interviewers, three video recordings were selected from batch 1 and three from batch 2 based on the availability of complete recordings as reported in Online Resource Table A.1. Two interviewers were males and five were females, with a median age of 30 years. All interviewers had a higher education, as three were graduating students and four were employed as researchers at a university. The respondents of the 42 selected interviews were aged between 18 and 66 years, with a mean age of 42 years. The average length of the interviews was 49 min.

The interview was identified as the minimum unit of analysis for the scoring and for the identification of improvements. The list of patterns was tested in a pilot round of scoring where six random interviews of interviewer 1 (i.e., three for batch 1 and three for batch 2) were scored up to the fifth non-practice cTTO task. This pilot was used to further refine the scoring system. The final version was used to score the remaining 36 video-recorded interviews. Two authors (CV and GA) independently scored all the selected video recordings except for their own interviews; a third scorer (MM) was involved in the scoring process to ensure impartiality and avoid any self-scoring bias. In this way, the total sample of 42 interviews was scored twice, but none of the interviewers scored their own interviews. The scoring was performed from the very beginning of the interview to the fifth TTO choice, as we believed that some of the checks could manifest in the early phases of the interview that preeceded the TTO tasks (e.g. "*the interviewer clearly illustrates the study purposes and tasks.*")

3 Results

3.1 Conversational Analysis

The conversational analysis identified a set of positive ($n = 20$) and negative ($n = 14$) interview patterns (Table 1). These patterns were classified into five sub-themes, which were further grouped into two main themes that referred to task execution and communication skills. The former

Table 1 Positive and negative patterns

ID	Item	Scale	Scale labels	References
PC01	The interviewer slows down a respondent attempting to task shortcutting/rushing	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the times	[20, 21]
PC02	The interviewer presents the two alternatives neutrally	Likert 1–5	1 = Never 2 = Rarely 3 = Sometimes–often 4 = Most of the time 5 = Always	
PC03	The interviewer explains that there are no right or wrong answers	Yes/No		
PC04	The interviewer clearly illustrates the study purposes and tasks	Likert 1–5	1 = Never 2 = Rarely 3 = Sometimes–often 4 = Most of the time 5 = Always	[21]
PC05	The interviewer supports the participant with useful information	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
PC06	Appropriate use of mouse	Likert 1–5	1 = Never 2 = Rarely 3 = Sometimes–often 4 = Most of the time 5 = Always	
PC07	The interviewer aids the respondent by referring to a previous example	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	[21]
PC08	The interviewer apologizes for any mistakes or technical troubles	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	[20]
PC09	The interviewer emphasizes or repeats the respondent's answer, to increase logical connection and coherence	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	[20]
PC10	The interviewer slows down in explaining the task, making it easy to understand	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
PC11	The interviewer corrects the participant and effectively addresses the misunderstanding	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
PC12	The interviewer is able to use multiple explanations for a certain issue	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
PC13	The interviewer provides useful syntheses of what has been done thus far	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
PC14	The interviewer copies the jargon of the respondent to make concepts easily understandable without affecting a correct explanation	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
PC15	The interviewer lets the respondent reflect before answering the questions	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
PC16	The interviewer encourages a respondent that is not very reactive	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
PC17	The interviewer appends questions to engage the respondent, or the respondent feels engaged enough to ask questions	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the times	[20]

Table 1 (continued)

ID	Item	Scale	Scale labels	References
PC18	The interviewer keeps the conversation's rhythm high	Likert 1–5	1 = Never 2 = Rarely 3 = Sometimes–often 4 = Most of the time 5 = Always	
PC19	The interviewer acknowledges that the respondent understood a question/congratulates him/her	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
PC20	The interviewer asks the respondent to read health states aloud	Likert 1–5	1 = Never 2 = Rarely 3 = Sometimes–often 4 = Most of the time 5 = Always	[21]
NC01	The interviewer provides an incorrect explanation of the task/tasks	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
NC02	The interviewer lets the respondent choose zero value without probing	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	[21]
NC03	The interviewer attempts to force an answer	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
NC04	The interviewer provides information that may influence the participant	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
NC05	The interviewer moves forward without being sure that the respondent fully understands	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
NC06	The interviewer skips some steps of the explanation or the interview process	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
NC07	The interviewer provides an incomplete explanation of the task	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
NC08	The interviewer provides a confusing explanation of the task	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
NC09	The interviewer provides redundant information	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
NC10	The interviewer interrupts the respondent at the wrong moment	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
NC11	The interviewer lets the respondent talk over him and loses the lead of the interview process	Likert 1–3	1 = Never 2 = Sometimes 3 = Most of the time	
NC12	The interviewer uses language that is too informal	Yes/No		
NC13	The interviewer's tone is too fast in explaining the task	Yes/No		
NC14	The interviewer's tone is tense, creating an unpleasant feeling	Yes/No		

NC negative check, *PC* positive check

includes positive/negative categories that refer to a correct/incorrect way to introduce and conduct the different parts and tasks of the interview, while the latter includes positive/negative categories that refer to the interviewer's positive and negative communication practices.

Online Resource Table A.3 reports themes and subthemes associated with each of the 34 patterns, and Fig. 1 reports a graph linking each item to its corresponding theme and sub-theme. The most frequent positive patterns were related to (1) correctness of interview execution, including
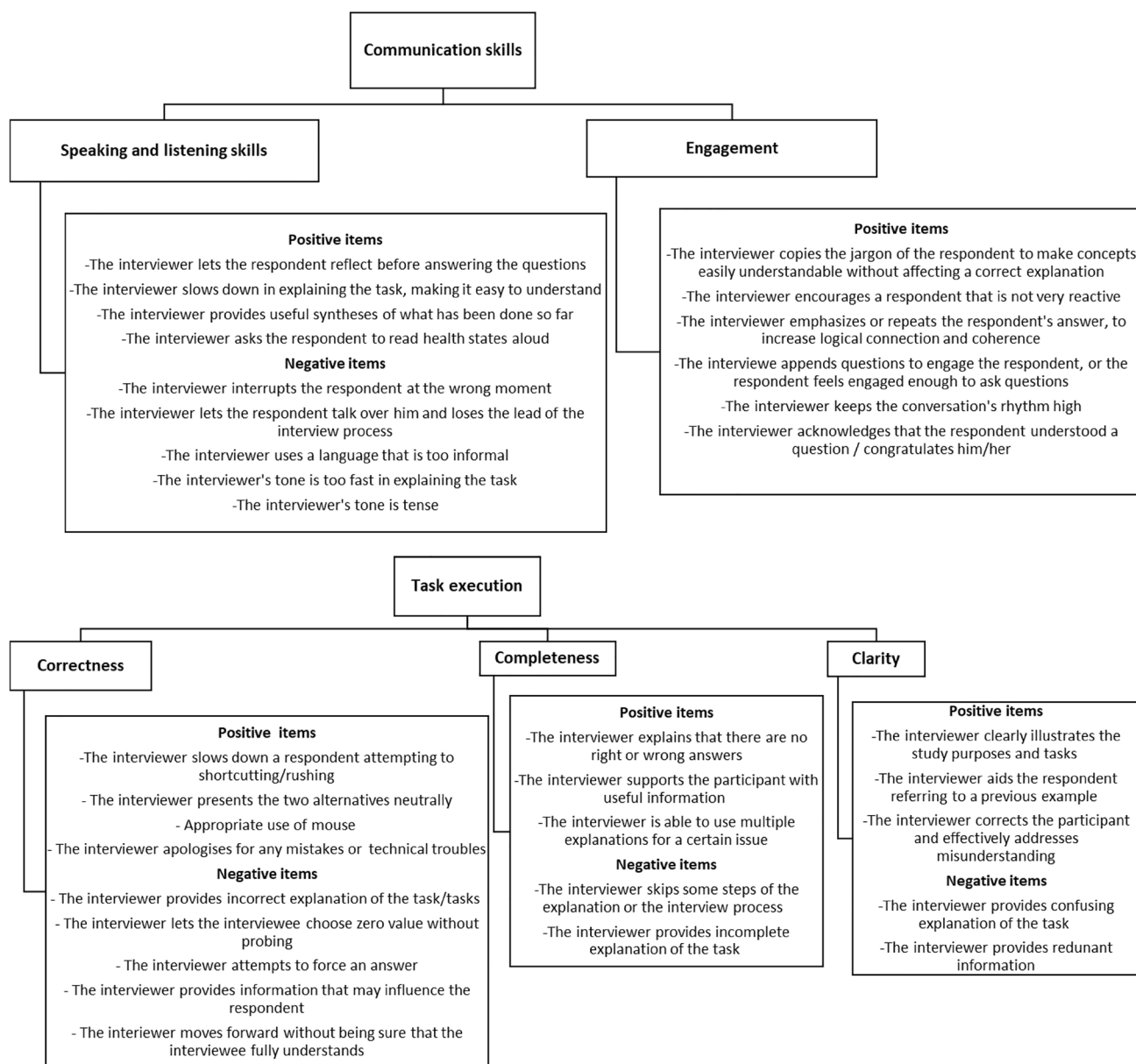
Fig. 1 Clusters of themes and sub-themes.

presenting alternatives neutrally and slowing down respondents who attempt to take shortcuts; (2) clarity, e.g., a clear explanation of the study purposes and tasks and using previous examples to enhance understanding; (3) speaking and listening skills, related to the tone of the speech and the turns between interviewer and respondent; and (4) mechanisms to attract the respondent's attention and motivate them to perform the task well.

The most frequent negative patterns were related to (1) speaking and listening skills, with the interviewers being too fast in their explanation or using incorrect wording and terminologies; and (2) correctness and completeness of task execution, e.g., not making sure that the respondent

understood the task or interrupting them in the wrong moment, or trying to force an answer or accepting an answer without making sure that the respondent understood the task. Table 2 presents some examples of how patterns from the conversational analysis were identified. Online Resource Table A.4 reports a list of the conversational analysis conventions used in the transcripts.

Table 2 presents a total of six examples, including two of positive patterns (Extracts 1–2) and four of negative patterns (Extracts 3–6).

- Extract 1 shows that, in two subsequent cTTO tasks, the interviewer (I) repeats several times the number

Table 2 Examples of positive and negative patterns

Extract 1 (P)	Interview 8 of batch 1 (TTO Tasks 2 and 3)	R: (0.4) In this case B (.) I: And would you give up maybe at 14? ↑ (0.3) or would you rather live, eeeh (0.5) i.e., you would rather live 10 years in full health and then 10 years this way (0.4) or is it better only 6 years of full health? ↑ (1) […] I: (0.4) And with respect to 5 years in full health? ↑ So possibly give up 5 years (0.8), would you be willing to give up 5 years of life to have them in full health (0.4) or would you still prefer to live longer (0.3) but in this state? ↑
Extract 2 (P)	Interview 1 of batch 1 (TTO Tasks 2 and 5)	I: Ok, (0.5) yes, (0.3) so this state in your opinion is more serious (0.3) than the previous one? (0.5) Do you remember the previous one? ↑ […] R: I would still prefer (.) the item A (.) (0.3) I: Item A (0.5), ok (.) (0.5). I always remind you to try to make a little comparison (0.5) with the initial one that we had done, (0.5) do you remember ↑ (0.5)? Where all these conditions were present but extreme (0.7). What we are interested in knowing is your opinion (0. 3) compared with that extreme state of health (0.5), where you had put that they were equal at this point.
Extract 3 (N)	Interview 2 of batch 3 (second trial state)	I: Full health, ok (0.3). And between dying now or living 10 years in this state (0.4)? Imagine for a moment what this could entail ... [R: [B, B (.) (0.3) I. In this case you would still prefer to live in this state (.) (0.8) R: Well, yes (0.3) I: Ok but between living 5 years in full health (0.3), so live half of the time ... [R: [A (.) (0.3) I: Ok (0.3), so you'd rather live half the time than in this state
Extract 4 (N)	Interview 1 of batch 1 (TTO task 3)	I: So, if you confirm it means that you would give up 20 years [R: [Sure (.) (0.3) I: In order not to experience (.) this condition in 10 years (.) ↓ [R: [Exactly (.)
Extract 5 (N)	Interview 7 of batch 5 (TTO task 2)	R: No, 4 years, no. (1) ° Can we put, for all of the state, more or less the same thing (0.3)?° I don't understand. [I: [Eh, let's say that the important thing for us is to have a nuance (0.3), the more serious the state, the more you should give up (0.5). <Maybe you can tell me that you would give up (0.3), if you consider it more serious than the other one, maybe... (0.8) > to have a gradient (0.3), to not have all the same ... [R: It goes well, here (0.3) < the fact is that you are seriously anxious and depressed > so that maybe you can bring problems to other people ... [I: [So we stop at four maybe ↑ (0.3) R: Okay four, yes (.)
Extract 6 (N)	Interview 7 of batch 2 (TT0 Tasks 2 and 3)	I: Would you prefer A then? (0.3) R: ° Yes ° (0.3) I: Ok [R: [Down (0.3), we arrive at a couple of years maximum ↓ (0.3) I: Ok (0.5), so would you give up = say 18 of 20 years? R: ° Mhmm hm. ° [I: [Ok perfect. [...] R: >I have no difficulty walking (0.3), I have no difficulty washing myself or dressing (0.5), I have no difficulty in carrying out my habitual activities, I am experiencing mild pain or discomfort (.) I am mildly anxious or depressed. < (0.5) A and B are equal, come on (0.3) I: You wouldn't give up even 6 months of life for this ... (0.7). Ok, so it's very slight let's say (0.3). Would you give up zero years out of 10? (0.3) R: Yes I: Ok (0.5) please

TTO time trade-off, *N* negative examples, *P* positive examples

of years to give up emphasizing the value, and/or s/he specifies the inverse of the trade-off for greater clarity (e.g., "give up to *x* years, then live only 10-*x* years?"). This is a way to ensure that the respondent (R) understands the task well and to increase the precision of the number of years to trade-off.

- In Extract 2, the interviewer shows a positive tendency of making references to previous examples to encourage

the respondent to think about the states previously evaluated and carefully weigh each level and dimension. This can be a successful way to prevent inconsistencies, e.g., when a health state is valued as worse than the 55555 state, which is the worst possible state deriving from the EQ-5D-5L descriptive system.

- In Extracts 3 and 4, the interviewer tries to read the cTTO alternatives but the respondent talks over him/her and the interviewer is not able to prevent this. This signals a limited grip on the interview process.

- In Extract 5 (first part), we observe that the respondent is going to select the same answer of the previous two states, where s/he had chosen to stop at 4½ years. Yet, the interviewer might be concerned about getting more diverse answers as per the QC guidelines, and so s/he 'cheats' and tries to force the respondent towards selecting the option of 4 years. In the second part, while the respondent is reasoning out loud before expressing his/her preference, the interviewer cuts him/her short, asking if s/he is fine with stopping at 4 years. This is likely to induce a bias towards the option suggested by the interviewer, as the respondent would probably have preferred to think more carefully about the state and s/he would have given a different answer.

- Extract 6 (first part) shows a respondent that speeds up to save time, as confirmed by the fact that a decision on the years to trade-off is made right after. Furthermore, the interviewer does not attempt to slow down a respondent who is short-cutting, and in doing so, s/he forgets an important step of the interview process. Similarly, in the second part of the extract, the respondent gets to zero immediately, stating the intention to give up his/her entire lifetime, and the interviewer fails to ask some prompting questions to try to curb a hastened response trend.

Online Resource Table A.5 reports an example for each of the patterns derived from the conversational analysis exercise, with the aim of giving examples of the language, speed and tone shades that we observed in each positive or negative behavior converging to what we termed 'patterns'. Patterns inspired from external sources [20, 21] or not verbal patterns, such as the appropriate use of mouse, are not included in the table.

3.2 Scoring the Video Recordings

The distribution of total normalized scores as given by the two independent scorers confirms that the discrepancies among scorers were relatively small. Thus, figures in this section will show the average of the two scorers. Online Resource Table A.2 provides the Delta table of the divergence of scores between two authors (CV and GA,

or MM where applicable). Figure 2 shows the distribution of positive and negative scores per interviewer across the six interviews scored for each. For most interviewers there is an improvement in positive patterns, but the size of this improvement differs across interviewers. Moreover, negative patterns decrease for most interviewers. In some cases, they increase but this is paired with an improvement in positive, meaning that the two even out, as is the case for interviewer 1. Instead, some interviewers show a rather stable trend in positive scores but a decrease in negative scores, as is the case for Interviewers 6 and 7. Interviewer 3 shows a stable increase in positive scores paired with a sharp decrease in negative scores.

Interestingly, for almost all interviewers ,we observed a sharp jump in the positive scores from the third interview to the fourth, namely from the last interview of Batch 1 to the first interview of Batch 2. Relatedly, for some interviewers, a consistent and decreasing trend in negative scores from the third interview (Batch 1, interview 3) to the fourth interview (Batch 2, interview 1) is observed. This might signal an improvement in positive interview practices following the verbal QC feedback that was provided to each interviewer after the second batch. The QC showed good performance of the interviewers already at baseline as there were limited to no issues in protocol compliance; however, some interviewers incurred interview-specific effects.

The standardized interview score for positive and negative outcomes is expressed on a 0–1 scale, with an average of 0.68 for positive checks and 0.49 for negative checks

Table 3 reports the average of positive and negative scores per interviewer and batch. Some interviewers, namely 3 and 7, showed both a decrease in the total score of negative items and an increase in the total scores of positive items (i.e., positive outcomes).

Other interviewers show mixed results, e.g., an improvement in the negative scores but a decrease in positive scores, or the other way around. A double negative outcome was only observed for Interviewer 1.

A full comparison of the results of this qualitative scoring exercise with the quantitative indicators resulting from the QC is beyond the scope of this study and remains a key avenue for further research. However, we highlight some interesting, although preliminary, results from the analysis of the main quantitative indicators of the QC, which are summarized in Table 4.

This table shows that interviewer 3, one of the two 'best performers' resulting from the qualitative scoring exercise, also shows better-than-average QC scores. Indeed, for this interviewer, we observe a duration above average for both the Worse than Death (WTD) and Better than Death (BTD) time trade-off (TTO) tasks, as well as a higher number of moves than average for both tasks.
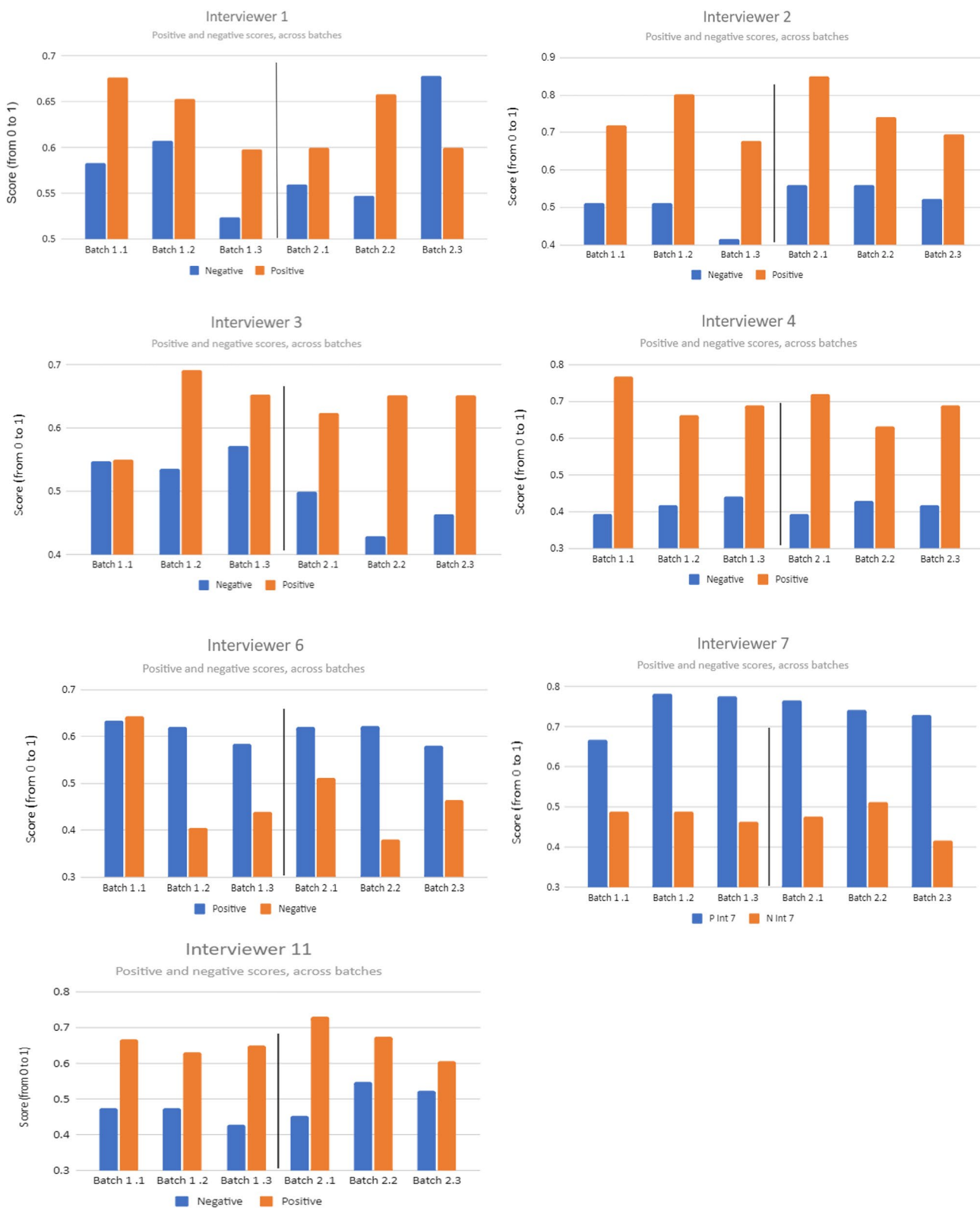
Fig. 2 Total positive and negative scores across the six interviews examined for each interviewer (average of scores from two scorers)

Table 3 Changes in overall scores from Batch 1 to Batch 2

Interviewer	Batch 1	Batch 2	Difference
Overall score for negative items			
Interviewer 01	0.571	0.595	0.024
Interviewer 02	0.480	0.548	0.067
Interviewer 03	0.552	0.464	**−0.087**
Interviewer 04	0.417	0.413	**−0.004**
Interviewer 06	0.496	0.452	**−0.044**
Interviewer 07	0.480	0.468	**−0.012**
Interviewer 11	0.460	0.508	0.048
Overall score for positive items			
Interviewer 01	0.643	0.619	−0.023
Interviewer 02	0.733	0.762	**0.029**
Interviewer 03	0.632	0.642	**0.011**
Interviewer 04	0.707	0.680	−0.027
Interviewer 06	0.613	0.608	−0.005
Interviewer 07	0.742	0.746	**0.003**
Interviewer 11	0.650	0.671	**0.021**

Cells highlighted in bold refer to improvements (decrease of negative scores or increase of positive scores); interviewer 3 and interviewer 7 showed both a decrease in negative scores and an increase in positive scores

Instead, Interviewer 1, the 'worst performer' of the qualitative scoring exercise, seems to perform below average for what pertains to the quantitative scores of the QC. For this interviewer, we observe a duration below average for both WTD and BTD TTO tasks and a lower number of moves than average for the TTO WTD task.

Finally, Table 5 reports those items for which there is a consistent improvement (decrease for negative and increase for positive). The only sub-theme whose items all improved from the first to the second batch was that of clarity.

4 Discussion

The QC procedure is a key component of the EQ-VT protocol. During the regular QC checks, the study team and EQ-VT support team evaluate protocol compliance and data distribution for all interviewers and face validity of the aggregate data. However, all the evidence considered in the QC process originates from quantitative indicators, and, thus far, there has been no investigation of the qualitative aspects of interviewers' performance. To fill in this gap, our study explored possible qualitative indicators of interviewers' performance in EQ-VT interviews by uncovering

Table 4 Changes in overall scores from Batch 1 to Batch 2

Interviewer	Batch	TTO time WTD	TTO time BTD	TTO moves WTD	TTO moves BTD
1	1	13.90	63.73	1.02	5.36
1	2	14.16	100.15	1.60	7.56
Average across batches		14.03	81.94	1.31	6.46
2	1	26.51	112.39	0.87	5.07
2	2	24.59	95.27	1.64	6.13
Average across batches		25.55	103.83	1.26	5.6
3	1	61.97	105.39	3.56	5.96
3	2	35.19	86.38	3.44	6.62
Average across batches		48.58	95.89	3.5	6.29
4	1	29.89	106.55	1.67	5.07
4	2	13.79	65.77	0.98	5.51
Average across batches		21.84	86.16	1.32	5.29
6	1	29.61	65.43	1.73	5.00
6	2	13.70	67.58	1.62	5.20
Average across batches		21.65	66.51	1.68	5.10
7	1	27.09	90.54	1.22	5.91
7	2	30.29	107.51	1.27	5.69
Average across batches		28.69	99.02	1.24	5.8
11	1	23.71	68.73	1.98	4.78
11	2	35.73	88.73	4.38	8.47
Average across batches		29.72	78.73	3.18	6.62
Average of all interviewers across all batches		27.15	87.44	1.93	5.88

TTO time trade-off, *WTD* Worse Than Death, *BTD* Better than Death

Table 5 List of items improving from Batch 1 to Batch 2

Category	Item	Sub-theme	Batch 1	Batch 2	Difference
N	NC01	Correctness	0.579	0.532	−0.048
N	NC05	Correctness	0.786	0.722	−0.063
N	NC07	Completeness	0.762	0.722	−0.040
N	NC08	Clarity	0.611	0.579	−0.032
N	NC09	Clarity	0.532	0.476	−0.056
N	NC11	Speaking and listening skills	0.738	0.667	−0.071
N	NC14	Speaking and listening skills	0.048	0.024	−0.024
P	PC02	Correctness	0.719	0.748	0.029
P	PC03	Completeness	0.929	1.000	0.071
P	PC04	Clarity	0.710	0.757	0.048
P	PC05	Completeness	0.857	0.905	0.048
P	PC06	Correctness	0.662	0.671	0.010
P	PC07	Clarity	0.563	0.595	0.032
P	PC11	Clarity	0.706	0.714	0.008
P	PC12	Completeness	0.690	0.730	0.040
P	PC15	Speaking and listening skills	0.849	0.873	0.024
P	PC16	Engagement	0.484	0.556	0.071
P	PC18	Engagement	0.710	0.771	0.062

N negative, *P* positive, *NC* negative check, *PC* positive check

conversational practices and patterns of interactions between the interviewer and the respondent.

Our study identified numerous positive and negative conversational patterns related to correctness, completeness, and clarity of interview execution, as well as communication skills. Examples of positive patterns included the ability of the interviewer to append questions that stimulated respondent's engagement, to provide different explanations for an aspect that was not clear, to support the participant with information that helped them complete the tasks, and to increase the interview coherence by confirming the respondent's answers. Examples of negative patterns included the interviewer moving forward in the exercise without making sure that the respondent had understood, the interviewer trying to force an answer from the respondent, and the interviewer speaking too fast or providing incomplete or incorrect explanations of the tasks.

Summarizing the qualitative findings of the conversational analysis, we developed a list of 20 positive and 14 negative patterns that allow us to assess interviewers' performance beyond what is measured in the QC reports. Using this list, we scored 42 randomly selected interviews from seven interviewers.

Although this study has an exploratory nature and the total number of scored interviews is somewhat limited to draw any firm conclusions regarding the usefulness of the QC, a few changes in patterns over time were identified that deserve discussion. It generally seems that the QC is having the expected effects, as witnessed by the consistency between the scores calculated in this study and the quantitative results of the QC report, and by the presence of numerous positive patterns that may result from the QC meetings and feedback provided to each interviewer. However, we identified some situations where respondents were being forced towards a certain answer—a behavior referred to as 'cheating', for which the interviewers' desire to comply with the QC can be one of the possible explanations. These can be instances of the unintended consequences of 'complying effects', as signaled by other researchers [10].

This is the first study to explore qualitative aspects of interviewers' performance in EQ-5D valuation studies. The positive and negative patterns identified may be useful to inform the training material of EQ-VT studies worldwide and complement the existing QC process, in particular to monitor interviewers' performance in case of suboptimal results distribution and high inconsistencies. We believe that the main contribution of this study is not the scoring system per se but the fact that it revealed several patterns, which were never discovered or presented before and which can aid in the interpretation and assessment of interviewers' performance.

Importantly, the study authors who analyzed the transcripts and then scored the interviews were shown the QC results of the related interviews only after the scoring phase had terminated. This had the purpose of avoiding potential bias: the over-identification of positive patterns and the under-identification of negative patterns in the transcripts of interviewers who obtained 'good' QC scores, and the opposite in the transcripts of interviewers who obtained 'bad' QC scores.

The present study also has limitations. The use of video-conferencing was made necessary by the outbreak of the coronavirus disease (COVID-19) pandemic. It should be noted that compared with in-person interviews, some subtle interactions could not emerge using this mode of administration [22], or the proportion of these patterns might be different from in-person settings. Moreover, we selected a subsample of interviews from two interviewers to inform the development of the list of patterns, therefore the list of positive and negative patterns we identified may not be exhaustive. As we did not score all the interviews in each batch, perfect matching with the QC indicators is not possible, and some effects observed may be due to random sampling. Moreover, for most interviewers, we only had interviews from the first and second batches of data collection. This means that we were not able to detect improvements that might have occurred at a later stage. In addition, non-verbal communication (e.g., facial expressions, hand gestures, voice, posture) was not taken into consideration, yet this might have great importance in conveying the message and it could be addressed in future research.

5 Conclusions

The patterns developed in this study can be easily applied to the recordings of valuation studies in other countries to improve their external validity and comparability. Moreover, a natural continuation of the present study would entail a comparison with the distribution of QC data and with the written feedbacks provided to each interviewer. In particular, the patterns deriving from the conversational analysis and average interviewer scores according to the newly developed scoring system may be assessed against (1) the average time for a cTTO task; (2) the clustering at 1, 0.5, 0, −0.5, and −1; (3) the proportion of inconsistent responses; (4) the mean utility value; and (5) the average number of moves in TTO. Thus, we believe that our approach has the potential to complement existing QC practices and provide useful insights for the improvement of future rounds of data collection in valuation studies worldwide and related EQ-VT indicators.

Declarations

References

1. Brooks R. EuroQol: the current state of play. Health Policy. 1996;37(1):53–72.
2. Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011;20(10):1727–36.
3. Feng YS, Kohlmann T, Janssen MF, Buchholz I. Psychometric properties of the EQ-5D-5L: a systematic review of the literature. Qual Life Res. 2021;30(3):647–73.
4. Finch AP, Brazier JE, Mukuria C, Bjorner JB. An exploratory study on using principal-component analysis and confirmatory factor analysis to identify bolt-on dimensions: the EQ-5D case study. Value Health. 2017;20(10):1362–75.
5. 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.
6. 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.
7. 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.

8. 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.

9. Dermawan Purba F, Hunfeld JAM, Iskandarsyah A, Sahidah Fitriana T, 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.

10. Viney R, Mulhern B, Norman R, Shah K, Devlin N (eds). Quality control vs. 'data curation': where should we draw the line in researcher judgements about the stated preference data used to value EQ-5D? In: 36th EuroQol Group Scientific Plenary Meeting in Brussels: Belgium, 18–21 Sep 2019.

11. Finch AP, Meregaglia M, Ciani O, Roudijk B, Jommi C. An EQ-5D-5L value set for Italy using videoconferencing interviews and feasibility of a new mode of administration. Soc Sci Med. 2022;292: 114519.

12. Meregaglia M, Malandrini F, Finch AP, Ciani O, Jommi C. EQ-5D-5L population norms for Italy. Appl Health Econ Health Policy. 2023;21(2):289–303.

13. 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(10):993–1004.

14. Stivers T, Sidnell J, editors. The handbook of conversation analysis. New Jersey: John Wiley & Sons; 2012.

15. Pallotti G. Conversation analysis: Methodology, machinery and application to specific settings. Conversation analysis and language for specific purposes; 2007. p. 37–67.

16. Wang NC. Understanding antibiotic overprescribing in China: a conversation analysis approach. Soc Sci Med. 2020;262: 113251.

17. Fatigante M, Heritage J, Alby F, Zuccermaglio C. Presenting treatment options in breast cancer consultations: advice and consent in Italian medical care. Soc Sci Med. 2020;266: 113175.

18. White AEC. When and how do surgeons initiate noticing of additional concerns? Soc Sci Med. 2020;244: 112320.

19. Schaeffer NC. Interaction before and during the survey interview: insights from conversation analysis. Int J Soc Res Methodol. 2021;24(2):181–202.

20. Arons A, Stolk E. Interviewer protocol adherence in the Dutch EQ-5D-5L valuation study. EuroQol Plenary meeting: Stockholm: 25–26 Sep 2014.

21. Garbarski D, Schaeffer NC, Dykema J. Interviewing practices, conversational practices, and rapport: responsiveness and engagement in the standardized survey interview. Sociol Methodol. 2016;46(1):1–38.

22. Rowen D, Mukuria C, Bray N, Carlton J, Longworth L, Meads D, et al. Assessing the comparative feasibility, acceptability and equivalence of videoconference interviews and face-to-face interviews using the time trade-off technique. Soc Sci Med. 2022;309: 115227.