Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

Quality Control Process for EQ-5D-5L Valuation Studies

Juan M. Ramos-Goñi, MSc[1,*], *Mark Oppe, PhD*[1], *Bernhard Slaap, PhD*[1], *Jan J.V. Busschbach, PhD*[2], *Elly Stolk, PhD*[1,3]

[1]*Executive Office, EuroQol Research Foundation, Rotterdam, The Netherlands;* [2]*Section of Medical Psychology, Department of Psychiatry, Erasmus MC, Rotterdam, The Netherlands;* [3]*Institute for Health Policy and Management, Erasmus University, Rotterdam, The Netherlands*

ABSTRACT

Background: The values of the five-level EuroQol five-dimensional questionnaire (EQ-5D-5L) are elicited using composite time trade-off and discrete choice experiments. Unfortunately, data quality issues and interviewer effects were observed in the first few EQ-5D-5L valuation studies. To prevent these issues from occurring in later studies, the EuroQol Group established a cyclic quality control (QC) process. **Objectives:** To describe this QC process and show its impact on data quality. **Methods:** A newly developed QC tool provided information about protocol compliance, interviewer effects, and mean values by health state severity. In a cyclic process, this information is initially used to evaluate whether new interviewers meet minimal quality requirements and later to provide feedback about how their performance may be improved. To investigate the impact of this cyclic process, we compared the quality of the data in Dutch and Spanish valuation studies that did not have this QC process with that in the follow-up studies in the same countries that used the QC process. Data quality was measured using protocol violations, variability between interviewers, the proportion of inconsistent responders, and clustering of composite time trade-off values. **Results:** In Spain, protocol violations were reduced from 87% in the valuation study to 5% in the follow-up study and in the Netherlands from 20% to 8%. In both countries, interviewers performed more homogeneously in the follow-up studies. The number of inconsistent respondents was reduced by 23.2% in Spain and 23.6% in the Netherlands. Values were less clustered in the follow-up studies. **Conclusions:** The implementation of a strict QC process in EQ-5D-5L valuation studies increases interviewer protocol compliance and promotes data quality.

Keywords: economic, health status index, life valuation, quality control, quality of life.

Introduction

The five-level EuroQol five-dimensional questionnaire (EQ-5D-5L) is a health-related quality-of-life instrument consisting of five dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), each with five levels of response (no problems, slight problems, moderate problems, severe problems, and extreme problems/unable) [1]. Instruments such as the EQ-5D are of great interest to clinical researchers and health economists to measure the benefit of health technologies. Two main reasons explain this interest. On one hand, the simplicity of the EQ-5D allows including it in any data collection process at a low burden for patients. On the other hand, the possibility to assign preference-based index values to the collected data makes it possible to use it in economic evaluation.

To develop a preference-based scoring algorithm, valuation studies to link EQ-5D-5L responses to index values are needed. To assess those values, the EuroQol Group developed a standardized protocol for such valuation studies [2,3]. It was implemented in a computer-assisted personal interview approach, called the EuroQol valuation technology (EQ-VT). The protocol centered around two valuation techniques: composite time trade-off (C-TTO) and discrete choice experiment (DCE). The C-TTO was developed and field-tested as part of a multinational research program [2,4]. It used the conventional TTO task for valuing health states considered better than death (BTD) [4,5], whereas it used lead-time TTO for health states considered worse than death (WTD) [4–9]. To promote comparability across EQ-5D-5L valuation studies, the interview was fully scripted and embedded in the EQ-VT. The script provided instructions about what standards and goals interviewers should achieve as well as text suggestions for what to say. The instructions section about how to explain C-TTO was notably detailed, anticipating the complexity of the C-TTO interviewer for both the respondent and the interviewer. Interviewers used an example health state ("being in a wheelchair") to explain the BTD and the WTD elements of the C-TTO task, by showing how the iterative procedure works, and interviewers are required to discuss the possibility that health states can be considered WTD.

[*] *Address correspondence to:* Juan M. Ramos-Goñi, Executive Office, EuroQol Research Foundation, Marten Meesweg 107, 3068 AV, Rotterdam, The Netherlands.

E-mail: jramos@euroqol.org.

http://dx.doi.org/10.1016/j.jval.2016.10.012

Spain and the Netherlands were among the first countries that used the EQ-VT for national valuation studies to obtain value sets for the EQ-5D-5L in 2012 and 2013, respectively. After the data were collected, preliminary analyses by both the Spanish and the Dutch research teams indicated interviewer effects: some interviewers systematically elicited higher values, lower values, or more inconsistent values than other interviewers [10,11]. This was not anticipated, because interviewer effects were not observed in a preceding pilot study that tested the application of the C-TTO technique [4]. A notable difference between the pilot and the national valuation studies was the experience of the interviewers with the C-TTO technique. The interviewers in the pilot study were researchers who participated in developing the interviewer instructions, whereas the interviewers in the Spanish and a large part of the interviewers in the Dutch valuation studies were inexperienced in conducting TTO experiments before participating in the EQ-VT studies. This led us to suspect that the observed interviewer effects could be caused by insufficient compliance with the protocol. With post hoc data cleaning it may be possible to mitigate biases in the resulting value set because of the presence of interviewer effects and data quality issues. Nevertheless, exclusion criteria are controversial and exclusions will reduce the sample size, which affects the power to estimate the value set and jeopardizes the representativeness of the sample of respondents in terms of background variables.

Acknowledging these difficulties, solutions were sought in tools to enhance protocol compliance so as to reduce interviewer effects and improve data quality. Such an approach was inspired by evidence regarding the benefit of quality control (QC) along randomized clinical trials [12–15]. Those QC processes are based on continuous data monitoring and various checks during data collection. Our specific case, however, is more challenging because we aimed to develop a standardized QC process that can be used in multiple countries, whereas preferences can vary across countries. Thus, it is not possible to distinguish between valid and invalid responses on the basis of values that might be obtained. We have developed a QC methodology and a software (EQ-VT QC tool) that does not place any previous assumptions on the values that might be obtained. We exploited the fact that the EQ-VT captures time and position of each mouse click, as well as the valuation data, which enables identification of possible patterns that emerge in valuation data in relation to key characteristics of each interviewer's approach to data collection. In this article, we describe the QC process for EQ-5D-5L valuation studies and explore the improvements in the resulting data, by comparing data from the first series of EQ-VT studies in Spain and the Netherlands, where the QC process was not available, with later data sets collected in the same countries, with the QC process in place.

Methods

QC Reports

The EQ-VT QC tool produced standardized reports including figures, tables, and the explanation of its content. In Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2016.10.012, you will find the full QC reports combined for both the Spanish studies. The instructions and handling of DCE are simpler than those of C-TTO and therefore the QC report focused more on the C-TTO than on the DCE. The QC report can be grouped into four main sections: sample demographic characteristics, assessment of protocol compliance, assessment of interviewer effects in the data, and assessment of the consistency of the data with respect to health state severity (measured as level sum scores). The first section is self-explanatory and the content of the latter three sections of the QC report is summarized in Table 1.

The QC report started presenting the number of interviews, the demographic profile of the sample, and the number of interviews per interviewer. This was followed by figures that related to protocol compliance, which are present per interviewer. For instance, the available timings and positions of the mouse clicks provide the time interviewers used for explaining the C-TTO task and how many moves in the iterative procedure the interviewers showed to the respondents. This duration was assessed separately for the BTD part and the WTD part of the C-TTO. Using this information, it was possible to determine whether interviewers violated the protocol. For instance, short durations of the whole interview or parts of the interview suggested that the interviewer rushed through the instructions. Furthermore, the position of the mouse clicks can show whether the interviewer omitted demonstrations of parts of the EQ-VT functionality, such as not showing the WTD task to the respondent.

The next section of the QC report reviewed the values per interviewer such as the proportion of nontraders (i.e., respondents who give all health states the value of 1.00), the proportion of zero values, the proportion of negative values, and the proportion of respondents who value the state 55555 better than at least another state. This part of the report also included a comparison of the mean value over all health states and the overall SD per interviewer. For all interviewers individually there were figures with the distribution of the values from the −1.00 to 1.00 utility scale. Ideally, these figures should show, and not much, differences between interviewers. The last section of the QC report focused on the assessment of the consistency of the pooled data over all interviewers with respect to health state severity.

For several items, criteria were chosen to distinguish between compliance and noncompliance. If items were in the range of noncompliance, an interview was given a "flag." The QC tool reported a table with the number and proportion of "flagged interviews" per interviewer. An interview was flagged if one of the following criteria was met:

1. the WTD element was not shown in the wheelchair example;
2. the time spent on explaining the C-TTO task in the wheelchair example was less than 3 minutes;
3. a respondent spent less than 5 minutes to complete the 10 C-TTO tasks; or
4. the value for state 55555 was not the lowest and it was at least 0.5 higher than that of the state with the lowest value.

The judgment about protocol compliance of the DCE task was operationalized by looking for suspicious response patterns. An example is that a respondent always chooses the health state on the left. The report included a summary of the suspicious DCE responses in a table. This table was organized as follows: interviewer (column 1); the number of interviews completed (column 2); the mean amount of time taken (in minutes) to complete the seven DCE tasks (column 3); and the number of respondents who used suspicious response patterns of choices across all seven DCE tasks (columns 4–7).

As mentioned earlier, quality issues seem less a problem for DCE because compliance with the DCE instructions seems easier. We did not determine a minimum standard at forehand to define compliance to the DCE protocol.

QC Process

Before the study, interviewers were trained: they had to conduct practice interviews before participating in the actual data collection. The practice interviews were reviewed with the QC tool and

Table 1 – QC report description.

Figure/table	Aim/expectations
Assessment of protocol compliance	
Interview total duration Amount of time taken to complete each C-TTO task Amount of time taken to complete each DCE task	The purpose of these three figures is to inform whether an interviewer systematically shortens tasks. Interviewers may want to perform fast interviews to finish their work earlier. We expect low variability between and within interviewer. In other words, all interviewers take similar time and all respondents take similar time for each interviewer.
Amount of time spent in the wheelchair example Time spent in the BTD element of the wheelchair example Time spent in the WTD element of the wheelchair example	The purpose of these three figures is to inform how long the interviewers take to explain the C-TTO task, because shortcutting in the explanation could influence the respondent to do so. We expect that all interviewers make similar explanation following the interviewer script, making mean times to be similar among interviewers, but we also expect that interviewers always make the same explanation, and so we expect low variability within interviewer. But some variability is expected within interviewer because of specific respondent questions or doubts.
Moves performed in the wheelchair example Moves performed in the BTD element of the wheelchair example Moves performed in the WTD element of the wheelchair example	The purpose of these three figures is to inform about how well the iterative procedure (the process to move up or down the number of full health years in the C-TTO task) was explained. Few moves could not explain how to reach the preferred respondents' responses. We expect a large number of moves across all interviewers on each interview. So we expect high means, but low variability within interviewer. As mentioned in the previous section, respondent questions could lead to more or less moves.
Percentage of interviews in which the WTD element of the wheelchair example was used	The purpose of this figure is to inform about whether interviewers explain or at least show the WTD element of the wheelchair example. We expect that interviewers always show the WTD when they introduce the C-TTO task. This is a key indicator for protocol compliance. If the WTD element of the C-TTO task is not explained, the WTD responses will be bias producing zero censor values.
Assessment of interviewer effects in the data	
Percentage of respondents whose TTO data contain at least one "inconsistency" in relation to health state 55555 Nontraders	The purpose of these two figures is to inform either about respondents' misunderstanding or their laziness. On one hand, lazy respondents could shorten the tasks by expressing their indifference point in the first step of the iterative procedure; if they do that for the 10 C-TTO tasks they are considered as nontraders. There could, however, be real nontraders as very religious respondent. On the other hand, valuing the state 55555 higher than other states could be a signal of task misunderstanding as the 55555 is the worst possible health state defined by the EQ-5D-5L. We expect few inconsistent/nontrader respondents. For example, many inconsistent respondents for a specific interviewer could mean poor task explanation, even when time and moves look appropriate.
Percentage of health states given a value of exactly 0 in the TTO tasks Percentage of health states given a value of <0 in the C-TTO tasks	The purpose of these two figures is to inform about possible issues with the WTD element of the C-TTO tasks. For example, many 0 values (spike at 0) with small number of negative values could indicate either that the interviewer is preventing WTD values or that the interviewer is not explaining well the WTD element of the C-TTO task. We expect similar results across all interviewers.
Mean and SD of C-TTO values Distribution of responses for each specific interviewer	The purpose of this set of figures is not only to identify whether interviewers are influencing respondents, but also to find the side to which responses are biased and the size of the bias. We expect that interviewers have similar mean and SD values, but also similar distribution of values if no bias is present. These distributions are challenging to assess, given the fact that we do not know a priori what the "correct" mean values and distributions should be. Therefore, these figures are interpreted by comparing the data from each interviewer to the pooled data from all interviewers. In this way, we can see which interviewers can be considered as outliers. This evaluation is also helpful to appraise to what extent the differences in interview style that become apparent from the protocol compliance section might affect the data.
Assessment of the consistency of the data with respect to health state severity	
Mean and SD of C-TTO values, by level sum score	The purpose of this figure is to inform about the logical basis of the results. For instance, an indication of low-quality data is observing low mean values for mild states or high values for severe states, because it could be a consequence of obtaining key values in the iterative procedure (spikes). We expect health states with lower level sum scores to have higher mean value than those with higher level sum scores. But we also expect the opposite for SD; in other

continued on next page

Table 1 – *continued*

Figure/table	Aim/expectations
	words, we expect more agreement in slight health states than in severe health states.
Overall C-TTO value distribution C-TTO value distribution, by level sum score	The purpose of these figures is to inform about possible spikes and gaps in range of values. Interviewers may be similar when they are compared against each other, but they could be all producing similar influence over respondents. With these figure we can prevent this fact. Expectations very much depend on the country, that is, on cultural/religion traditions etc.

Note. 1) See Supplemental Materials found at http://dx.doi.org/10.1016/j.jval.2016.10.012 for examples of full reports; 2) When the within variability of the aggregate data for one interviewer is too high compared with others, an outlier is present.
BTD, better than death; C-TTO, composite time trade-off; DCE, discrete choice experiments; EQ-5D-5L, five-level EuroQol five-dimensional questionnaire; QC, quality control; WTD, worse than death.

interviewers were given feedback on their performance. The QC reports from the first 10 interviews done per each interviewer were used to evaluate whether they met the minimum quality requirements to contribute to data collection. When a new interviewer had conducted 10 interviews and 4 or more were flagged, all 10 interviews conducted by that interviewer were removed from the database and he or she required retraining. After a further 10 interviews, the performance was re-evaluated. If again 4 or more interviews were flagged, these interviews were also removed and the interviewer was removed from the interviewer team. The 40% threshold was selected over more stringent cutoff points because the criteria listed earlier do not capture interviewer performance perfectly; sometimes respondents might be responsible for flags. Only when problems seemed persistent, interviewer performance was considered to be the main problem. In later stages of the data collection, the QC reports allowed the study teams to reflect on interviewer's performance and gave them continuous feedback about how to improve. This cyclic nature of the process provided a continuous stream of information that allowed the interviewers to keep improving their skills during the entire data collection period.

Data

We used data from the Spanish and Dutch EQ-5D-5L valuation studies [10,11] that were conducted without the QC process, and data from two follow-up studies in the same countries in which the QC process was implemented [16,17].

The Spanish and Dutch valuation studies followed the EQ-5D-5L valuation protocol as described by Oppe et al. [3], now known as version 1.0 of the protocol. The protocol had three main sections. In the first section, interviewers explained the purpose of the study and respondents were requested to value their own health using the EQ-5D-5L and were asked about their background characteristics. The second section of the interview consisted of the C-TTO tasks. This started with the interviewer explaining the C-TTO task using the example of being in a wheelchair as the health state. After this explanation, the respondents were asked to value 10 EQ-5D-5L health states using C-TTO. The last section of the interview consisted of a DCE in which respondents were requested to answer seven paired comparisons, each consisting of two EQ-5D-5L states. The two follow-up studies used an updated version of the protocol, known as version 1.1, which included the same C-TTO and DCE tasks as version 1.0. Nevertheless, several improvements were implemented: 1) three practice states (mild: 21121; severe: 35554; and moderate but difficult to imagine: 15411) were added immediately after the wheelchair example to better prepare respondents for the C-TTO task; 2) respondents were offered the possibility to confirm their response before starting the next task; and 3) the cyclic QC process was implemented as described in previous sections. The same interviewer instructions were used in both versions, except for the added instructions about the three practice states in version 1.1. All interviews for all studies were performed face-to-face using the EQ-VT platform.

The two follow-up studies were part of a methodological research program that was launched to address the data quality issues that were reported in the first wave of the EQ-5-5L valuation studies (version 1.0). These follow-up studies compared data collected using version 1.1 of the protocol with data collected using experimental versions of protocol 1.1, during

Table 2 – Protocol compliance.

Variable	Spain		The Netherlands	
	Valuation study	Follow-up study	Valuation study	Follow-up study
Sample size	89	196	107	205
Proportion of interviews flagged (N)	87% (77)	5% (10)	20% (21)	8% (17)
Proportion of interviews in which the WTD element was not used in the wheelchair example (N)	71% (63)	0.5% (1)	9% (10)	1% (3)
Proportion of interviews in which interviewer did not spend at least 180 s (3 min) on the wheelchair example (N)	76% (68)	0.5% (1)	5% (5)	2% (4)
Proportion of interviews in which interviewer did not spend 5 min on the 10 TTO tasks (N)	34% (30)	4% (7)	3% (3)	2% (5)

TTO, time trade-off; WTD, worse than death.

which further modifications of the protocol were tested. For the present assessment of the QC process, we use data collected with only version 1.1 of the protocol in the follow-up studies to avoid confounding with other changes to the protocol.

Health States

The protocol for EQ-5D-5L valuation studies included 86 health states in the design of the C-TTO task. Only 10 of those states were also used in the follow-up studies. We restricted the comparison of the C-TTO data generated by the different versions of the protocol to those 10 states: 12111, 11122, 42321, 13224, 35311, 34232, 52335, 24445, 43555, and 55555. The DCE design was the same for all studies and included 196 health states distributed over 28 blocks of seven pairs of states.

Data Collection

C-TTO and DCE responses used for the comparison were derived from 597 participants included in the analysis, of whom 89 (Spain) and 107 (the Netherlands) participated in the valuation studies and 196 (Spain) and 205 (the Netherlands) participated in the follow-up studies. Respondents from the valuation studies were recruited from a panel using quota sampling, and those from the follow-up studies using convenient samples.

Interviews in the Dutch and Spanish valuation studies were conducted by 21 and 32 trained interviewers, respectively. The training in Spain consisted of one half-day session covering the interviewer instructions. In the Netherlands, the training took a whole day and consisted of a presentation of the components of the interview, discussion of the interviewer instructions, practice interviews in pairs, and a discussion of difficult interview elements. In addition, the Dutch principal investigator reached out at least once to each interviewer after data collection had started to discuss the interviewer's experiences.

For the follow-up studies, six trained interviewers in the Netherlands and seven in Spain, different from the valuation studies in both countries, conducted all interviews during March to April 2014. The training in Spain now consisted of a 3-day workshop, covering study background and aim, interviewer script, 10 practice interviews for each interviewer, plus a round table to share/comment interview issues and to review the QC reports for the 10 practice interviews. The training in the Netherlands involved the same 1-day training session as the valuation study. In both follow-up studies, interviewers were monitored at least weekly using the EQ-VT QC tool, which described the quality of their interviews.

Analysis

For between-study comparisons, we used proportions to present protocol compliance results. Means, standard errors, and variation coefficients (SD/mean) of duration and number of moves in the wheelchair example for both BTD and WTD values were used to explore the harmonization level of the C-TTO explanation within each study. We compared the interviewer effects by using a graphical presentation of kernel distributions of values for each interviewer. In addition, we compared the overall distribution of values and the distribution of values for state 55555 to illustrate values of the most severe health state. Finally, we considered the proportion of inconsistent respondents. In this analysis, an inconsistent respondent is defined as a respondent who values at least one pair of logically dominant health states inconsistently. We used the Paretian Classification of Health Change [18] in the definition of a logical dominance relationship between two health states; that is, state 1 dominates state 2 when state 1 is better than state 2 on at least one dimension, and no worse than state 2 on any remaining dimension. Therefore, the value of state 1 should be higher than the value of state 2, and when it is lower we considered it as an inconsistency.

Results

The results showed that protocol compliance was an issue in the Spanish valuation study: 87% of interviews had a protocol violation. For example, the interviewers did not explain the WTD element of the wheelchair example in 71% of interviews. In contrast, the Spanish follow-up study had 5% of protocol violations and 0.5% of interviews omitting the WTD in the wheelchair example (Table 2). In the Dutch valuation study, protocol compliance was less of an issue, with interviewers violating the protocol in 20% of cases. Nevertheless, improvements were still made because the proportion of protocol violations dropped to 8% in their follow-up study. Various indicators were also affected. For example, the average time taken to explain the C-TTO task using the wheelchair example, the average time per TTO task, and the number of moves used in the iterative procedure all increased. In addition, the variation

Table 3 – Wheelchair example (duration and number of moves).

	Spain				The Netherlands			
Indicator	Valuation study		Follow-up study		Valuation study		Follow-up study	
	Mean (SE)	Variation coefficient	Mean (SE)	Variation coefficient	Mean (SE)	Variation coefficient	Mean (SE)	Variation coefficient
Total time (s)	132 (14)	1.21	661 (15)	0.32	433 (15)	0.35	368 (8)	0.33
Time on BTD element (s)	110 (13)	1.10	446 (11)	0.35	297 (11)	0.38	241 (6)	0.36
Time on WTD element (s)	22 (7)	3.09	215 (7)	0.44	136 (8)	0.60	126 (4)	0.49
Total moves	9.1 (1.0)	1.01	42.3 (1.0)	0.33	25.8 (1.2)	0.49	24.6 (0.8)	0.44
Moves on BTD element	7.3 (0.9)	1.10	25.5 (0.7)	0.38	17.6 (1.0)	0.56	15.3 (0.5)	0.50
Moves on WTD element	1.7 (0.4)	2.30	16.8 (0.6)	0.48	8.1 (0.7)	0.88	9.3 (0.4)	0.64

BTD, better than death; SE, standard error; WTD, worse than death.

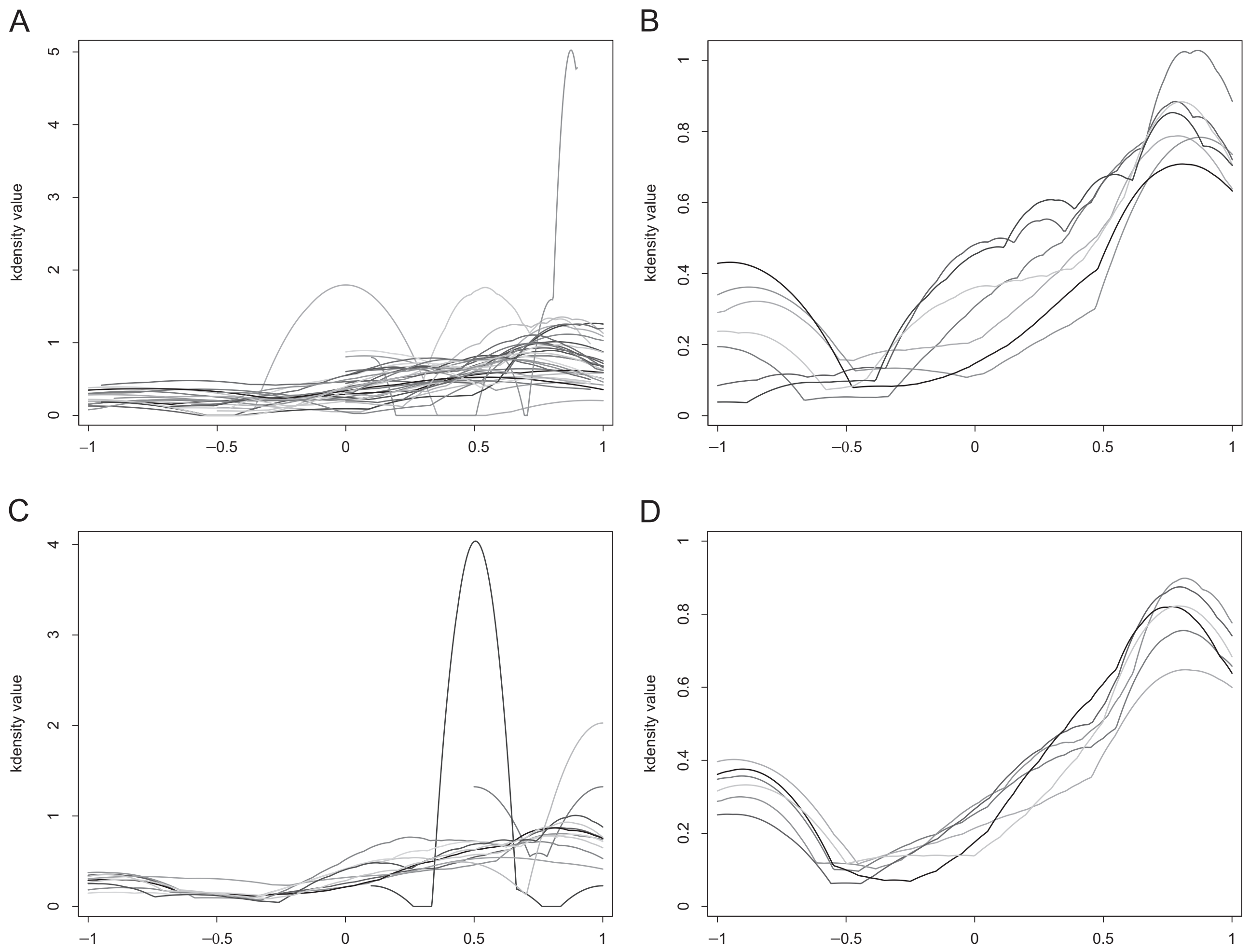

Fig. 1 – Distribution of values over the 10 health states by interviewer.

coefficients were smaller for durations and moves spent on the C-TTO explanation in the follow-up studies compared with valuation studies: they all showed more homogeneity (Table 3).

With respect to the interviewer effects, the distribution of values per interviewer was more homogeneous in the follow-up studies compared with that in the valuation studies in both countries (Fig. 1). In particular, the Spanish valuation study showed 11 of the 32 interviewers eliciting no WTD values, whereas all interviewers did in the follow-up study (Fig. 1A,B); further details are provided in the Supplemental Materials. The Dutch valuation study showed 1 out of 21 interviewers eliciting no WTD values, whereas all interviewers did in the follow-up study (Fig. 1C,D). Both the overall distribution and the distribution for the state 55555 showed less clustering of values at 0 in the follow-up studies. The proportion of negative values was higher in the follow-up study, lowering the mean observed value for state 55555 in both countries. The gap of values between −0.5 and 0 shown in the valuation studies is mitigated in the follow-up studies (Fig. 2).

The QC process also had an impact on the proportion of respondents with one or more inconsistent responses. The proportion of Spanish respondents who had at least one inconsistent response was 48.3% for the valuation study, whereas this proportion dropped to 25.1% in the follow-up study. In the Netherlands, these proportions were 43.9% and 19.5% for the valuation and the follow-up study, respectively. Differences in proportions of inconsistent respondents were significant ($P < 0.0001$).

Discussion

This article reported on the effect of implementing a QC process on EQ-5D-5L valuation studies, using four data sets, two of which were collected with QC and two without QC. The results provide evidence that our QC process improved the quality of the valuation data, but the effect size of the improvements varied between countries because of the marked difference in the data quality obtained in their valuation studies without the QC process. In this study, we assumed that the increased time taken for the interviews and the increased number of moves in the iterative C-TTO task reflect greater interviewer and respondent engagement. Moreover, the reduction in inconsistent responses and a lower clustering of values were also seen as improvements. One can argue that the extensive training and QC influenced interviewer responses more. This might be true, but it is difficult to see that why inconsistent and clustered responses represent better answers. All in all, the QC process seemed to improve the data in a valuation study, whereas uncertainty exists about the quality of captured data if QC is not adopted.

The QC process presented here, although custom-made for EQ-5D-5L valuation studies, was built on the same principles as those of the traditional QC process to check units of production [19]. There are, however, obvious limitations in our case. Our QC process was more challenging, because our unit of production was a set of subjective values elicited in an interview so that neither the validity of the values nor the validity of the interview can be directly

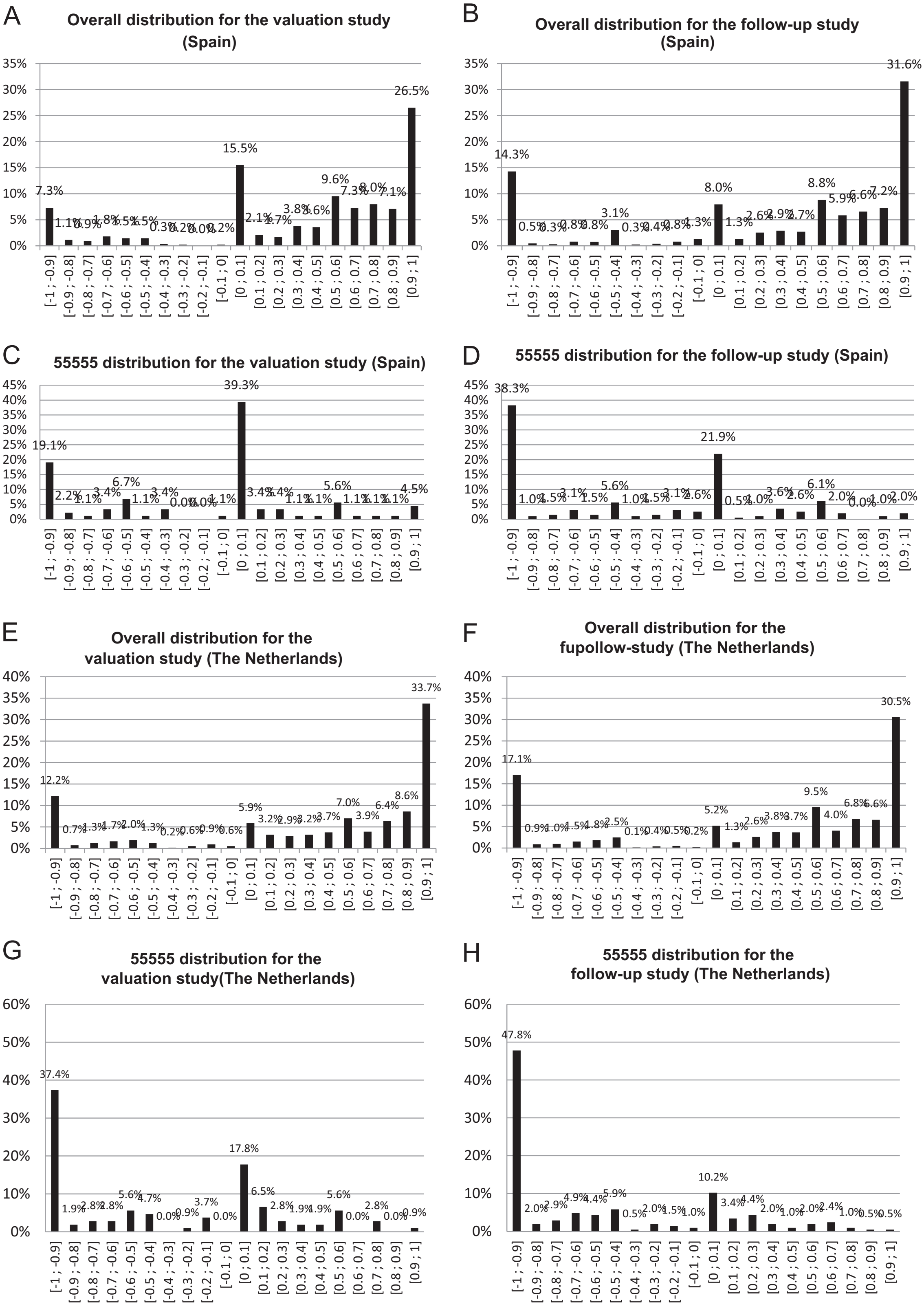

Fig. 2 – Overall and pits state (55555) distributions.

appraised. Each interview was unique making arbitrary the definition of a valid interview. It was unrealistic to expect that each interviewer will use exactly the same wording in all interviews, or as his or her colleagues; it depended on the questions from the respondents. This led us to focus on averages and variability, rather than using interviews as units, making it possible to harmonize interviewer performance and reduce potential bias in respondents' responses. The effect of the QC process then partly came via the continuous monitoring and feedback process instigated by the principal investigator, and was not simply a result of taking out the bad units. Rather, to account for respondent interviewer interaction, we established a conservative threshold of 4 flagged interviews out of 10 as the limit to stop and retrain the interviewer. This was our analogy process of stopping and reviewing our production system. The information about the appropriate actions to take when issues are encountered can be found elsewhere [20].

Study Limitations

The QC process is probably not the only factor that caused the observed differences between valuation and follow-up studies. Small modifications beyond the QC process were also introduced from version 1.0 used in valuation studies to version 1.1 used in follow-up studies, such as the introduction of practice states and a confirmatory pop-up screen. Nevertheless, it is unlikely that these additions to the protocol can substantially improve interviewer compliance with the protocol, which was achieved by monitoring the time interviewers spent on each part of the interview, as part of the QC process.

Arguably, the improvement in training efforts in Spain may have had an important impact on the data as well, impeding conclusions about causality or about to what extent improvements can be attributed to the QC process. Nevertheless, because we also observed an improvement in the Netherlands, where the initial training efforts were comparable across the two studies, it is reasonable to assume that at least part of the effect size can be attributed to the QC. Further studies should clarify which parts of the QC process are most helpful. Another limitation of this study is that differences between respondents may affect results. It should, however, affect only the distribution of values by interviewer comparison, but it should not affect neither the overall distribution nor the overall inconsistency rate.

Conclusions

The results in this article support the decision of the EuroQol Group to extend its original EQ-5D-5L valuation protocol 1.0 with a QC tool. The impact of the QC process on the characteristics of a data set is large. We therefore recommend an uptake of similar strategies in future valuation studies, that is, transparency about interviewers' selection and training and the kind of feedback they received about their performance. Key characteristics of the raw data need to be reported as well to make possible judgment about the quality. Past valuation studies, including most of the EQ-5D-3L valuation studies, lack this kind of transparency. It is likely that efforts to prevent data quality issues across valuation studies will help to improve the determination of difference that related to cultural, methodological, analytical, or procedural choices. The implementation of the cyclic QC on EQ-5D-5L valuation studies increased interviewer protocol compliance, reduced differences in an interviewer's elicited values, and significantly improved data quality.

Acknowledgments

Most of the research effort is funded by the EuroQol Research Foundation. Additional funds came from the Academic Centers. All authors are associate with the EuroQol Group. We express our gratitude to Koonal Shah for the texts he provided to accompany the tables and figures in the QC reports. We also express their gratitude to Oliver Rivero-Arias for the useful feedback while preparing this article.

Source of financial support: This work was partially funded by the EuroQol Group.

Supplemental Materials

Supplemental material accompanying this article can be found in the online version as a hyperlink at http://dx.doi.org/10.1016/j.jval.2016.10.012 or, if a hard copy of article, at www.valueinhealthjournal.com/issues (select volume, issue, and article).

REFERENCES

[1] Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011;20:1727–36.

[2] Devlin N, Krabbe P. The development of new research methods for the valuation of EQ-5D-5L. Eur J Health Econ 2013;14(Suppl.):1–3.

[3] Oppe M, Devlin NJ, van Hout B, et al. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health 2014;17:445–53.

[4] Janssen BM, Oppe M, Versteegh MM, et al. Introducing the composite time trade-off: a test of feasibility and face validity. Eur J Health Econ 2013;14(Suppl. 1):S5–13.

[5] Oppe M, Rand-Hendriksen K, Shah K, et al. EuroQol protocols for time trade-off valuation of health outcomes. Pharmacoeconomics 2016;34:993–1004.

[6] Robinson A, Spencer A. Exploring challenges to TTO utilities: valuing states worse than dead. Health Econ 2006;15:393–402.

[7] Augustovski F, Rey-Ares L, Irazola V, et al. Lead versus lag-time trade-off variants: does it make any difference? Eur J Health Econ 2013;14 (Suppl. 1):S25–31.

[8] Luo N, Li M, Stolk EA, et al. The effects of lead time and visual aids in TTO valuation: a study of the EQ-VT framework. Eur J Health Econ 2013;14(Suppl. 1):S15–24.

[9] Devlin N, Buckingham K, Shah K, Tsuchiya A, Tilling C, Wilkinson G, vanHout B. A comparison of alternative variants of the lead and lag time TTO. Health Econ. 2013 May;22(5):517–32.

[10] Ramos-Goñi JM, Pinto-Prades JL, Oppe M, et al. Valuation and modeling of EQ-5D-5L health states using a hybrid approach. Med Care (published online ahead of print December 17, 2014). http://dx.doi.org/10.1097/QAI.0000000000000492.

[11] Versteegh MM, Vermeulen KM, Prenger R, et al. Dutch tariff for the 5 level version of EQ-5D. Value Health 2016;19(4):343–52.

[12] Knatterud GL. Methods of quality control and of continuous audit procedures for controlled clinical trials. Control Clin Trials 1981;1:327–32.

[13] Vantongelen K, Rotmensz N, van der Schueren E. Quality control of validity of data collected in clinical trials. EORTC Study Group on Data Management (SGDM). Eur J Cancer Clin Oncol 1989;25:1241–7.

[14] De Pauw M. Quality control in data monitoring of clinical trials. Acta Urol Belg 1994;62:31–5.

[15] Buyse M. Centralized statistical monitoring as a way to improve the quality of clinical data. March 24, 2014. Available from: http://www.appliedclinicaltrialsonline.com/centralized-statistical-monitoring-way-improve-quality-clinical-data. [Accessed June 10, 2016].

[16] Ramos-Goñi JM, Rand-Hendriksen K, Pinto-Prades JL. Reintroduction of the ranking task in valuation studies: improved data quality and reduced level of inconsistencies? The case for EQ-5D-5L. Value Health. 2016;19(4):478-86.

[17] Shah K, Rand-Hendriksen K, Ramos-Goñi JM, Prause AJ, Stolk E. Improving the quality of data collected in EQ-5D-5L valuation studies: a summary of the EQ-VT research methodology programme. 31ST EuroQol Group Scientific Plenary. Stockholm, Sweden. Sep 2014.

[18] Devlin NJ, Parkin D, Browne J. Patient-reported outcome measures in the NHS: new methods for analysing and reporting EQ-5D data. Health Econ 2010;19:886–905.

[19] Radford GS. The Control of Quality in Manufacturing. New York, NY: Ronald Press Co., 1922.

[20] Purba FD, Hunfeld JA, Iskandarsyah A, Fitriana TS, Sadarjoen SS, Passchier J, Busschbach JJ. Employing quality control and feedback to the EQ-5D-5L valuation protocol to improve the quality of data collection. Qual Life Res. 2016 Oct 31. [Epub ahead of print].