QC report for England

EuroQol Office,

Rotterdam, The Netherlands

March 2019

Note. This report was generated after all data were collected not during data collection. The English data was collected in 2012-2013 using EQVT v 1.0. The QC tool used to generate this report was first implemented in EQVT v 1.1. released in 2014.

QC report for England: Executive summary

The English EQ-5D-5L study was conducted in 2012-2013, in the first wave of EQ-5D-5L valuation studies. A set of problems encountered in the cTTO data of this first wave of valuation studies led to the subsequent refinement of the valuation protocol. The updated version of the protocol elicit values using the same valuation techniques, but the way of implementing the tasks has improved. Based on the recognition that accuracy of collected responses and level of detail provided by interviewers in their explanation of the task varied, EuroQol introduced a quality control (QC) procedure to review protocol compliance and interviewer effects. A post hoc analysis of data quality is reported here to examine interviewer performance in the English EQ-5D-5L study.

The key findings of the QC report are presented on page 3 (table 2, reporting protocol compliance) and page 13 (figure 18, depicting interviewer effects in collected responses). The picture that emerges from the data is as follows:

1. Interviewer Compliance with protocol was low (page 5, table 2).

The team would have kept only 138 of the 998 interviews if they had followed the current guidance. The remaining 860 interviews would not be part of the final dataset, because the low levels of protocol compliance would have prompted early intervention. For every interviewer who failed to meet the QC requirements initially (i.e. after 10 interviews), this batch of 10 interviews would have been discarded and the interviewer would have been retrained. They would only be allowed to contribute new data if improved performance was confirmed.

2. Strong interviewer effects are present in the data (page 21, fig. 16).

Figure 18 shows little similarities in the distribution of collected responses across interviewers, suggesting interviewer effects. The particular way of an interviewer to ask questions or accept an answer has influenced the responses. Strong clustering is found in the responses from some interviewers. This suggests censoring/satisficing/lack of task understanding as discussed by Stolk et al (2019). Interviewer and clustering effects usually dissolve over time as interviewer performance improves with experience and received feedback. In England that has not happened because the fieldwork was completed by 50 interviewers who completed on average 20 interviews. This means that little benefit could be incurred from the feedback offered by the study team.

These findings need to be interpreted with care. Presence of protocol violations does not automatically imply bad data because it is defined on procedural grounds. For

instance, protocol is violated if the worse than dead task is not explained in the wheelchair example (applicable to 70% of the interviews), but it is possible that the explanation has been given later, when a respondent was confronted for the first time with a health state worse than dead. Yet, the opposite is also true. If a respondent did not value any state as worse-than-dead, one would like to confirm that the respondent was aware of the possibility to give negative values in order to distinguish censored observations from responses that reflect true values.

Similarly, interviewer effects and clustering of data show that collected responses have low accuracy. Depending on treatment such observations receive when they are modelled, the resulting values may nevertheless be found to reflect true preferences well. The careful way in which the English team searched for behavioural explanations for patterns in the data to inform their modelling approaches suggests higher levels of external validity than one might expect considering only face validity of the data.

QC Report for EQVT study

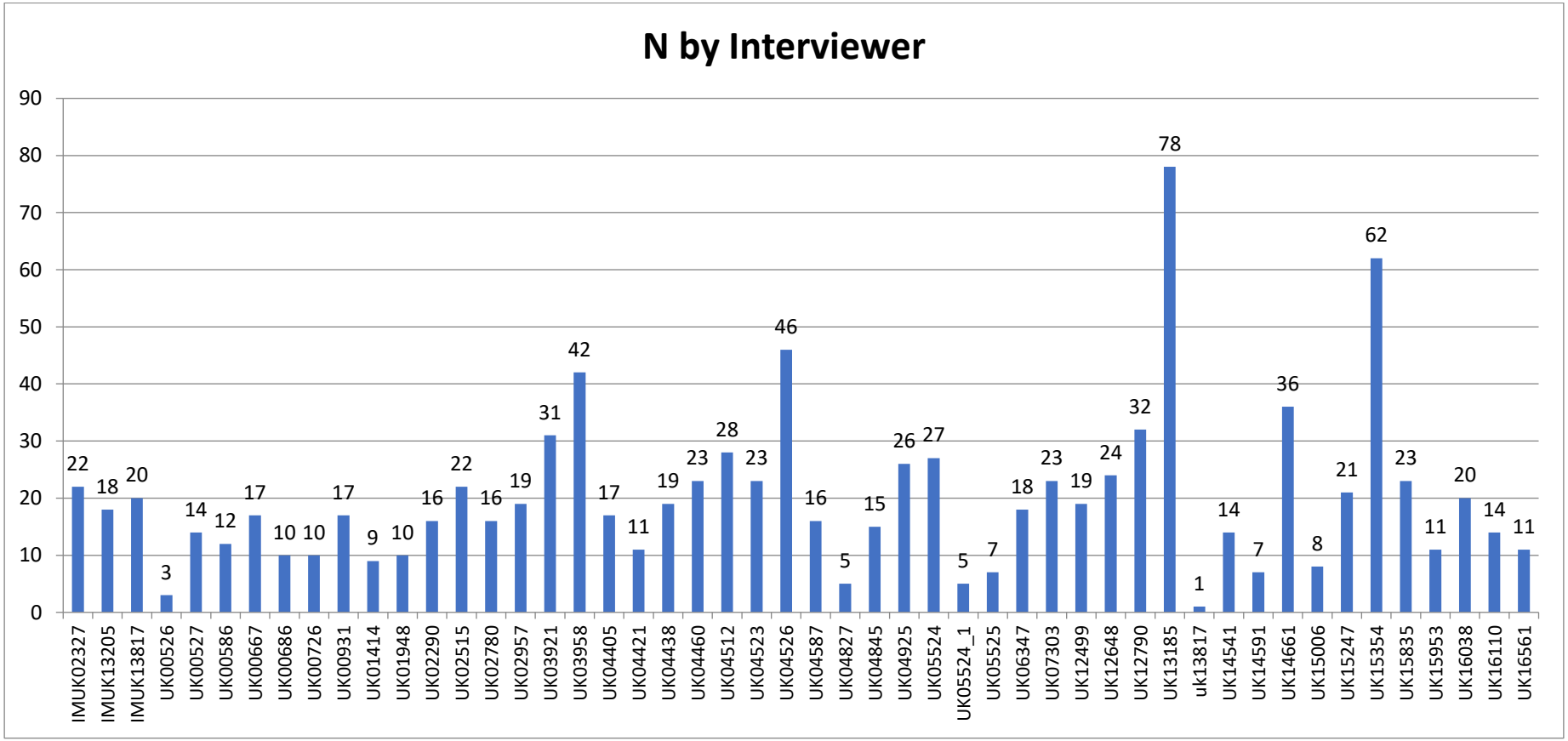
This document is automatically generated by the EuroQol EQ-VT QC Excel tool.

Date of report: 06/03/2019

Table 1. Sample representativeness

Age Range	FEMALE	MALE	Total
<25	31	19	**50**
[25 - 34]	97	55	**152**
[35 - 44]	123	86	**209**
[45 - 54]	95	78	**173**
[55 - 64]	88	68	**156**
[65 - 74]	78	52	**130**
>75	81	47	**128**
Total	**593**	**405**	**998**

Figure 1. Number of interviews by interviewer



In EQVT v 1.0 no minimum or maximum number of interviews per interviewer was defined. Currently we require that every interviewer completes at least 70 interviews. The reasoning behind this is that the TTO interview is complex and all interviewers will be on a learning curve. After every round of data collection, interviewers receive feedback about their performance, to promote protocol compliance and reduce interviewer effects. Little benefits can be incurred from a feedback mechanism like this if each interviewer performs few interviews as was the case in England.

Protocol compliance, by interviewer

Table 2. Flagged interviews

Interviewer	N	N flagged	% flagged	WC LT	% WC LT	Incon size	% Incon size	WC time	% WC time	TTO time	% TTO time
IMUK02327	22	4	18%	2	9%	1	5%	2	9%	1	5%
IMUK13205	18	18	100%	17	94%	1	6%	18	100%	1	6%
IMUK13817	20	20	100%	15	75%	2	10%	19	95%	0	0%
UK00526	3	3	100%	3	100%	1	33%	2	67%	1	33%
UK00527	14	14	100%	11	79%	1	7%	6	43%	3	21%
UK00586	12	6	50%	1	8%	4	33%	2	17%	3	25%
UK00667	17	17	100%	16	94%	2	12%	14	82%	2	12%
UK00686	10	5	50%	2	20%	2	20%	3	30%	0	0%
UK00726	10	10	100%	9	90%	1	10%	9	90%	3	30%
UK00931	17	3	18%	0	0%	1	6%	2	12%	0	0%
UK01414	9	7	78%	6	67%	0	0%	1	11%	0	0%
UK01948	10	8	80%	8	80%	0	0%	5	50%	2	20%
UK02290	16	15	94%	7	44%	1	6%	14	88%	1	6%
UK02515	22	22	100%	20	91%	5	23%	0	0%	1	5%
UK02780	16	15	94%	12	75%	1	6%	14	88%	5	31%
UK02957	19	18	95%	17	89%	2	11%	17	89%	3	16%
UK03921	31	28	90%	27	87%	2	6%	21	68%	2	6%
UK03958	42	40	95%	26	62%	2	5%	32	76%	5	12%
UK04405	17	3	18%	0	0%	3	18%	0	0%	0	0%
UK04421	11	3	27%	3	27%	0	0%	1	9%	0	0%
UK04438	19	18	95%	18	95%	1	5%	16	84%	1	5%
UK04460	23	16	70%	8	35%	1	4%	13	57%	2	9%
UK04512	28	23	82%	11	39%	3	11%	18	64%	6	21%
UK04523	23	7	30%	0	0%	4	17%	3	13%	1	4%
UK04526	46	46	100%	45	98%	10	22%	46	100%	12	26%
UK04587	16	14	88%	14	88%	1	6%	10	63%	1	6%
UK04827	5	5	100%	4	80%	0	0%	5	100%	1	20%
UK04845	15	4	27%	2	13%	1	7%	0	0%	1	7%
UK04925	26	26	100%	26	100%	0	0%	16	62%	2	8%
UK05524	27	24	89%	21	78%	2	7%	6	22%	2	7%

Interviewer	N	N flagged	% flagged	WC LT	% WC LT	Incon size	% Incon size	WC time	% WC time	TTO time	% TTO time
UK05524_1	5	5	100%	5	100%	0	0%	2	40%	0	0%
UK05525	7	4	57%	3	43%	1	14%	0	0%	0	0%
UK06347	18	8	44%	6	33%	2	11%	1	6%	1	6%
UK07303	23	23	100%	23	100%	0	0%	14	61%	4	17%
UK12499	19	19	100%	17	89%	1	5%	16	84%	6	32%
UK12648	24	22	92%	19	79%	2	8%	18	75%	4	17%
UK12790	32	29	91%	23	72%	5	16%	25	78%	0	0%
UK13185	78	78	100%	68	87%	10	13%	77	99%	60	77%
uk13817	1	0	0%	0	0%	0	0%	0	0%	0	0%
UK14541	14	12	86%	12	86%	1	7%	5	36%	0	0%
UK14591	7	6	86%	6	86%	0	0%	1	14%	0	0%
UK14661	36	17	47%	4	11%	2	6%	14	39%	4	11%
UK15006	8	8	100%	6	75%	0	0%	8	100%	3	38%
UK15247	21	20	95%	15	71%	3	14%	17	81%	8	38%
UK15354	62	38	61%	1	2%	5	8%	36	58%	3	5%
UK15835	23	23	100%	18	78%	5	22%	23	100%	12	52%
UK15953	11	10	91%	8	73%	0	0%	8	73%	3	27%
UK16038	20	10	50%	6	30%	2	10%	4	20%	1	5%
UK16110	14	4	29%	1	7%	0	0%	3	21%	1	7%
UK16561	11	10	91%	10	91%	0	0%	2	18%	0	0%

This table shows how many times each interviewer's TTO data have been flagged for data quality reasons. The total number of flagged interviews is shown in column 2, and the proportion of flagged interviews is shown in column 3. A given interview may be flagged for more than one reason. The flags are defined below. Table 2 suggests low levels of protocol compliance.

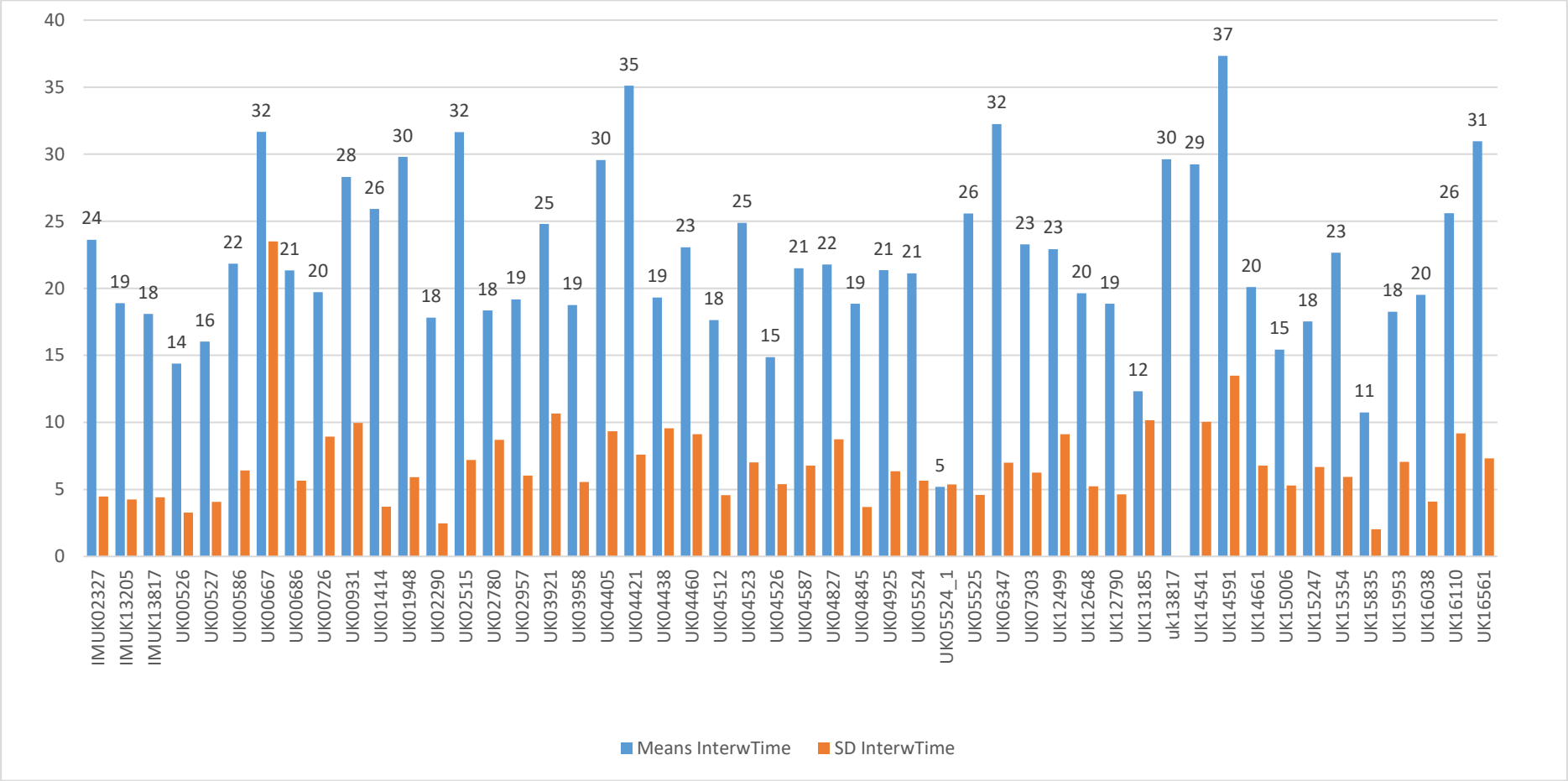
1) WC LT - Interview is flagged if the interviewer does not explain the worse-than-dead element in the wheelchair example.

2) Incon size - Interview is flagged if the respondent has a clear inconsistency in their TTO ratings (the value for 55555 is not the lowest and is at least 0.5 higher than that of the state with the lowest value).

3) WC time - Interview is flagged if the interviewer does not spend at least 180 seconds (3 minutes) on the wheelchair example.

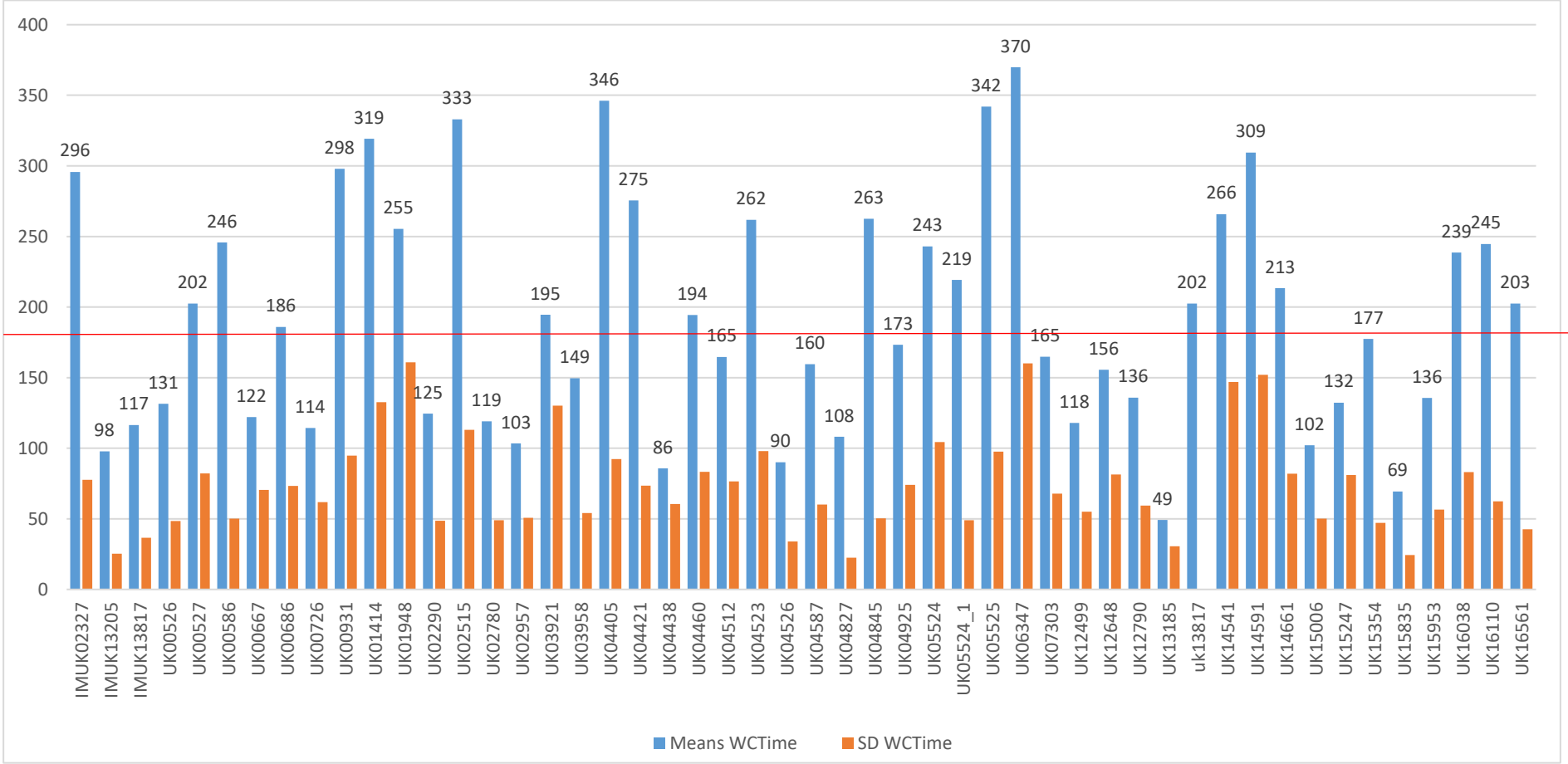
4) TTO time - Interview is flagged if the respondent does not spend at least 5 minutes on the 10 TTO tasks

Figure 2. Duration of interviews, by interviewer



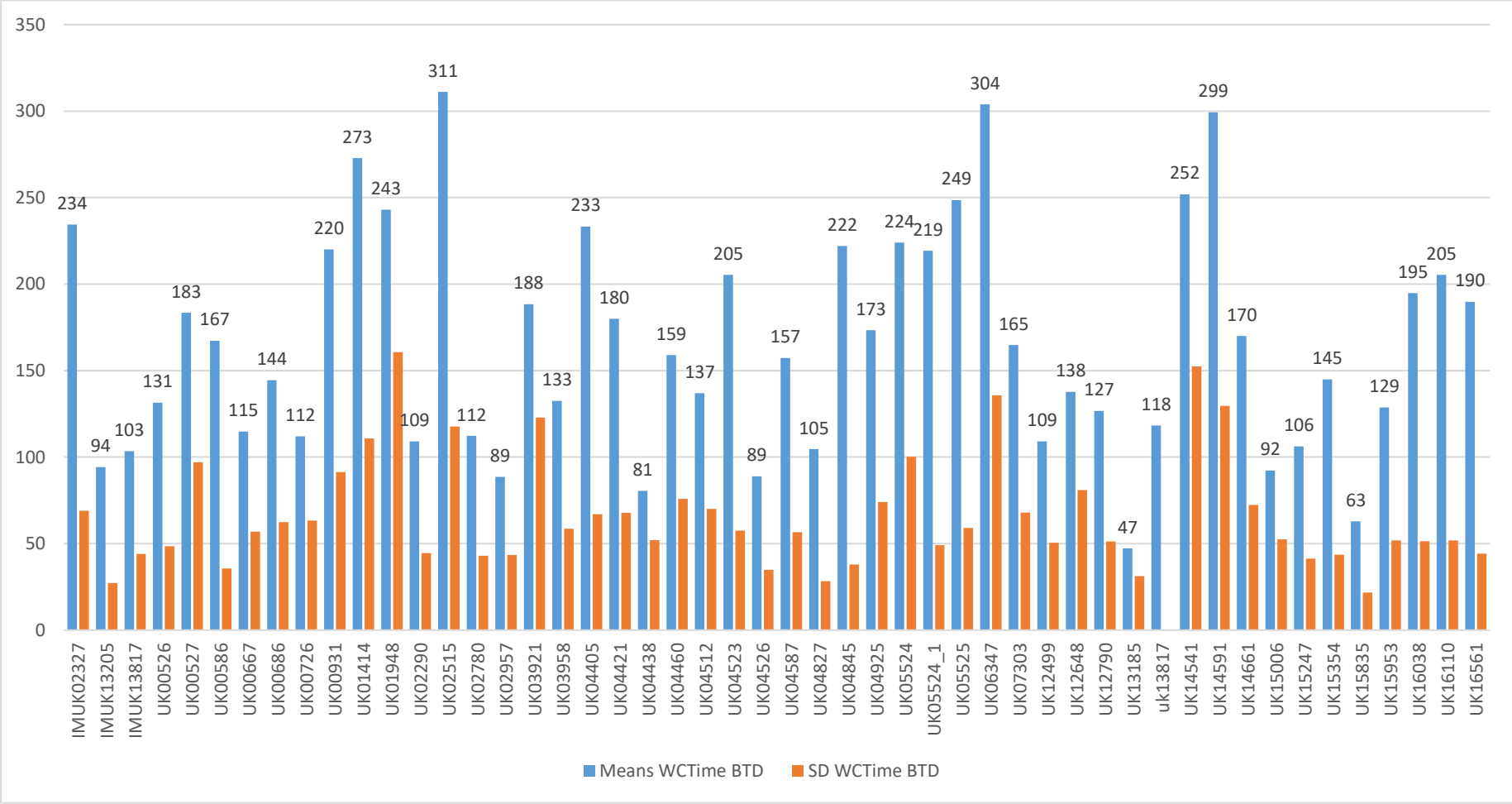
This figure shows the mean (and standard deviation) amount of time taken (in minutes) to complete the valuation questionnaire, by interviewer. This excludes any time taken to complete additional questionnaires such as the country-specific background questionnaire.

Figure 3. Time spent on TTO wheelchair example, by interviewer



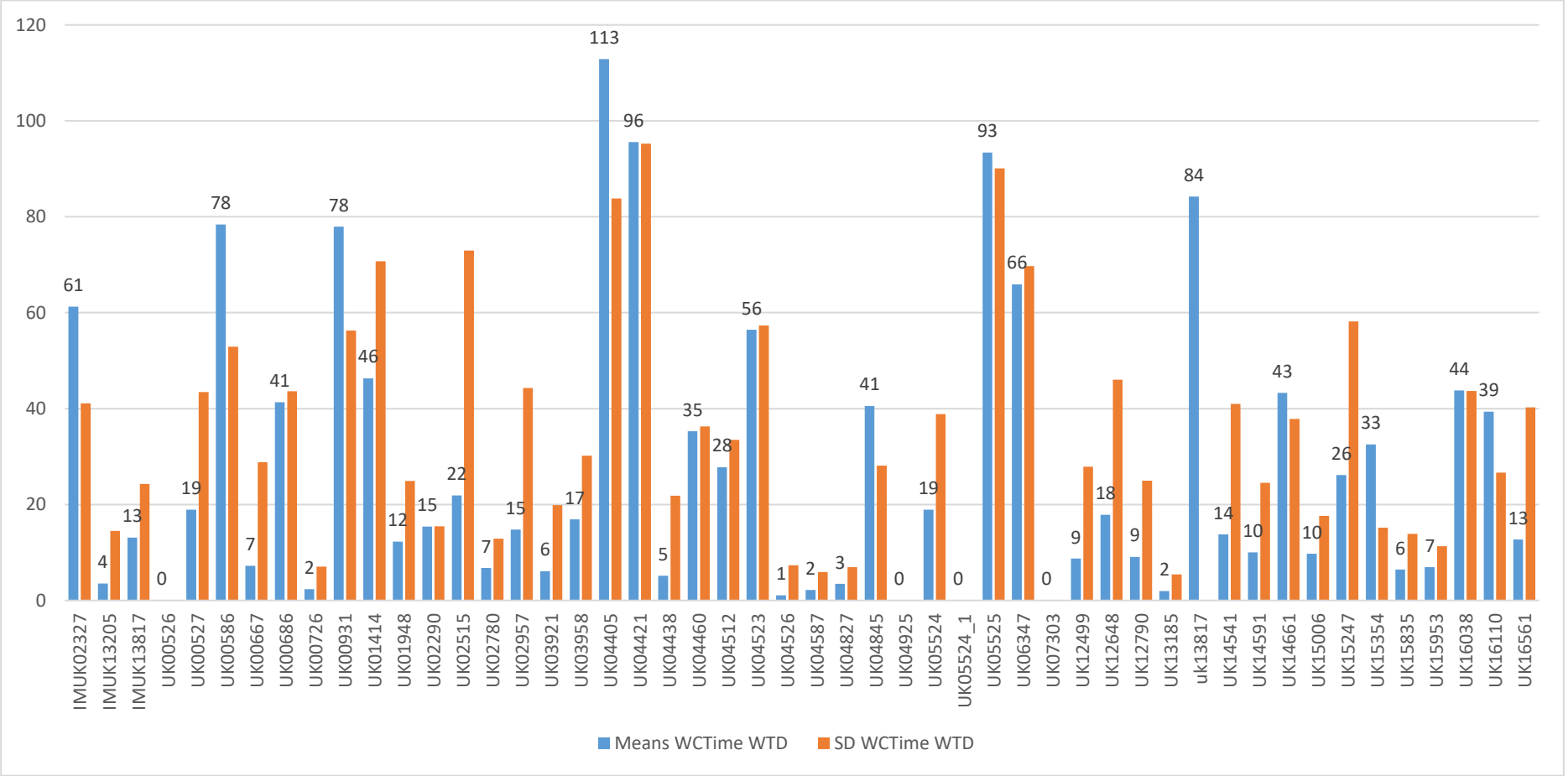
This figure shows the mean (and standard deviation) time spent (in seconds) on the wheelchair example, by interviewer. Interviewers have to explain all aspects of the TTO tasks during this example and those explanations are fully scripted. If someone explains the task in less than 180 seconds, this is considered a protocol violation because this implies that (most likely) several parts of the explanation have been omitted. Figure 4-9 provide detail on the use of the wheelchair example by interviewer.

Figure 4. Time spent on BTD element of TTO wheelchair example, by interviewer



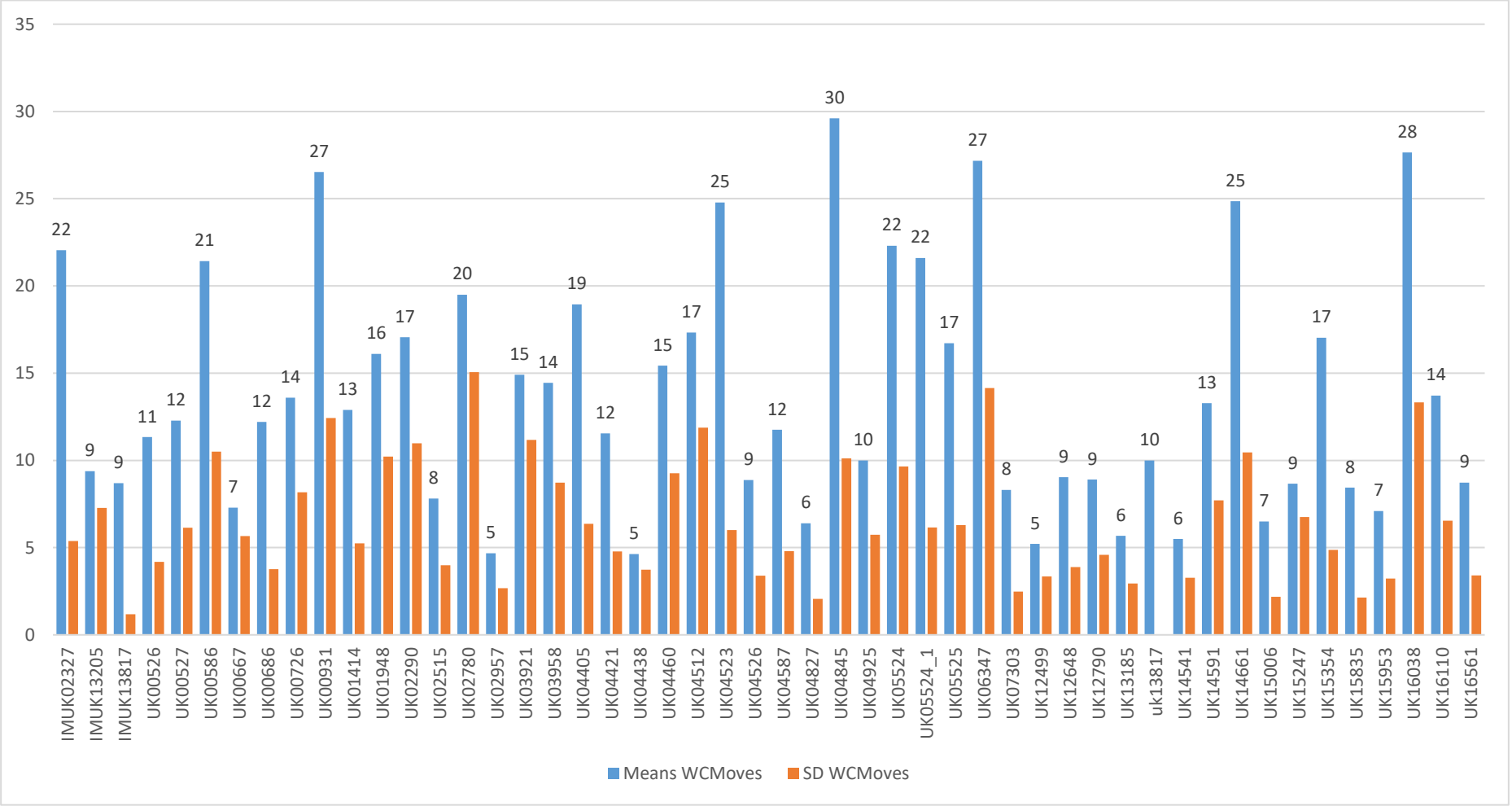
This figure shows the mean (and standard deviation) amount of time spent (in seconds) on the better-than-dead element of the wheelchair example, by interviewer.

Figure 5. Time spent on WTD element of TTO wheelchair example, by interviewer



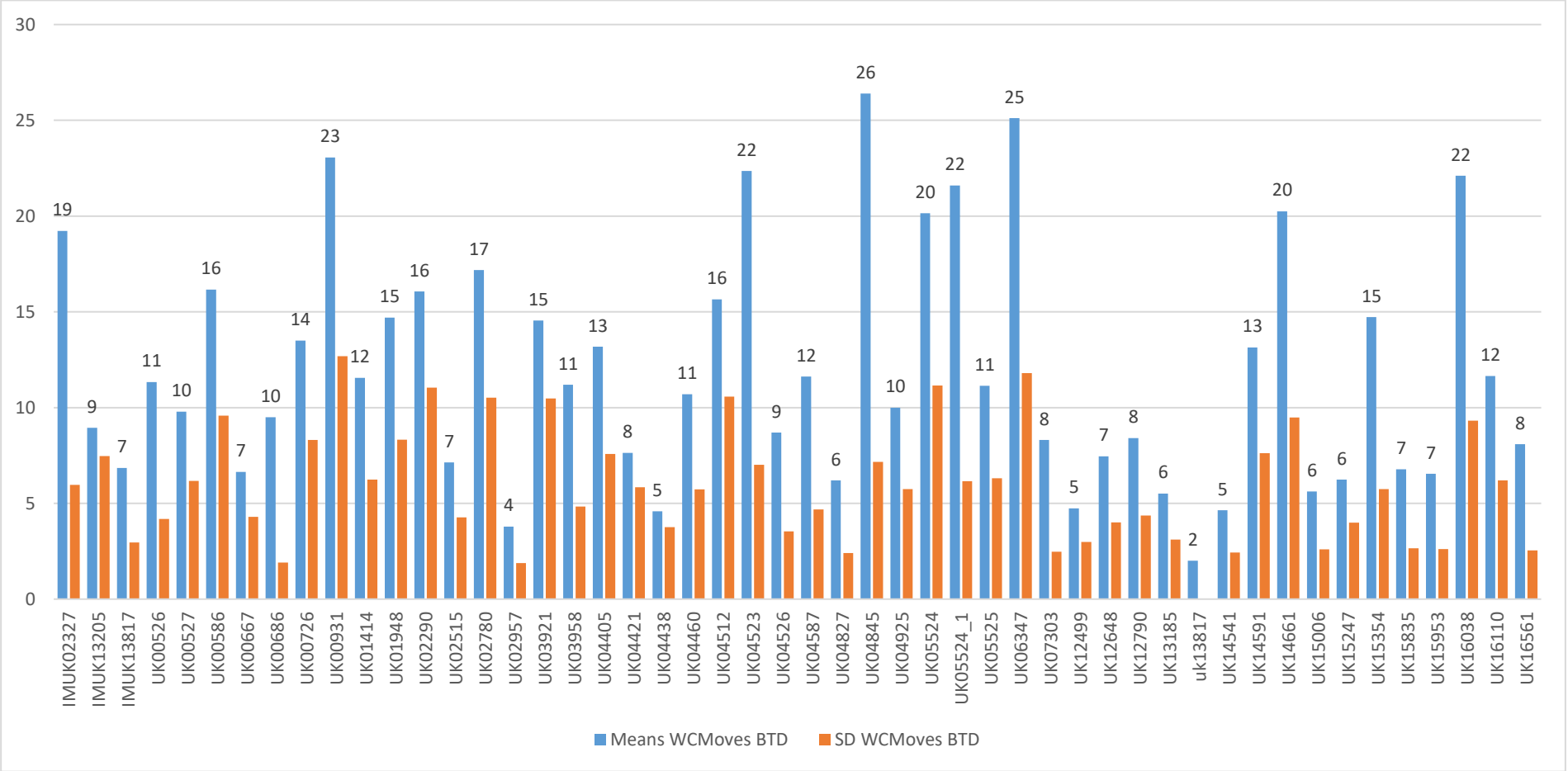
This figure shows the mean (and standard deviation) amount of time spent (in seconds) on the worse-than-dead element of the wheelchair example (designed to introduce the lead time TTO task), by interviewer.

Figure 6. Number of moves used to complete the TTO wheelchair example, by interviewer



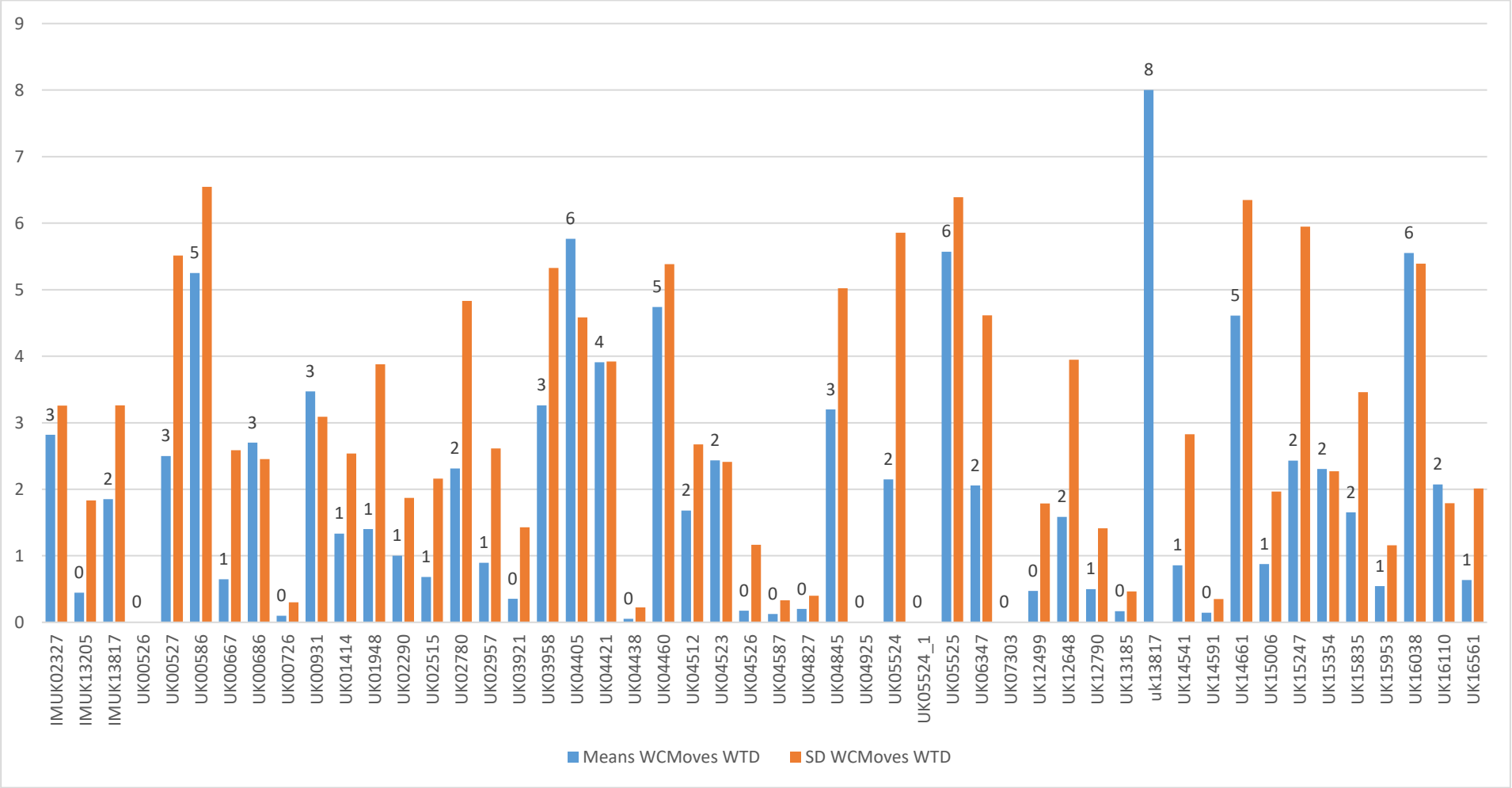
This figure shows the mean (and standard deviation) number of iterative steps used in the wheelchair example, by interviewer.

Figure 7. Number of moves used in BTD element of the TTO wheelchair example, by interviewer



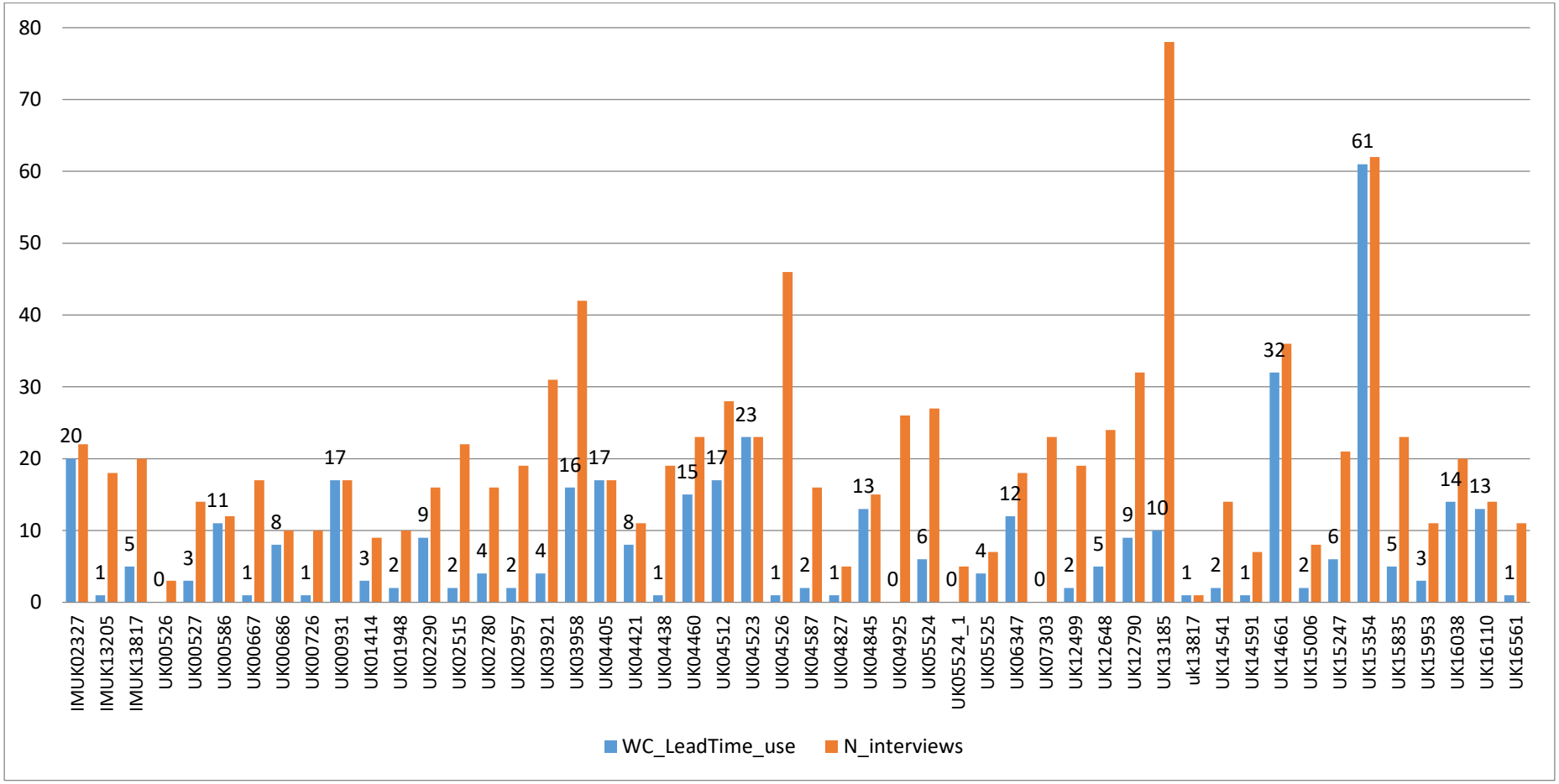
This figure shows the mean (and standard deviation) number of iterative steps used in the better-than-dead element of the wheelchair example, by interviewer.

Figure 8. Number of moves used in WTD element of the TTO wheelchair example, by interviewer



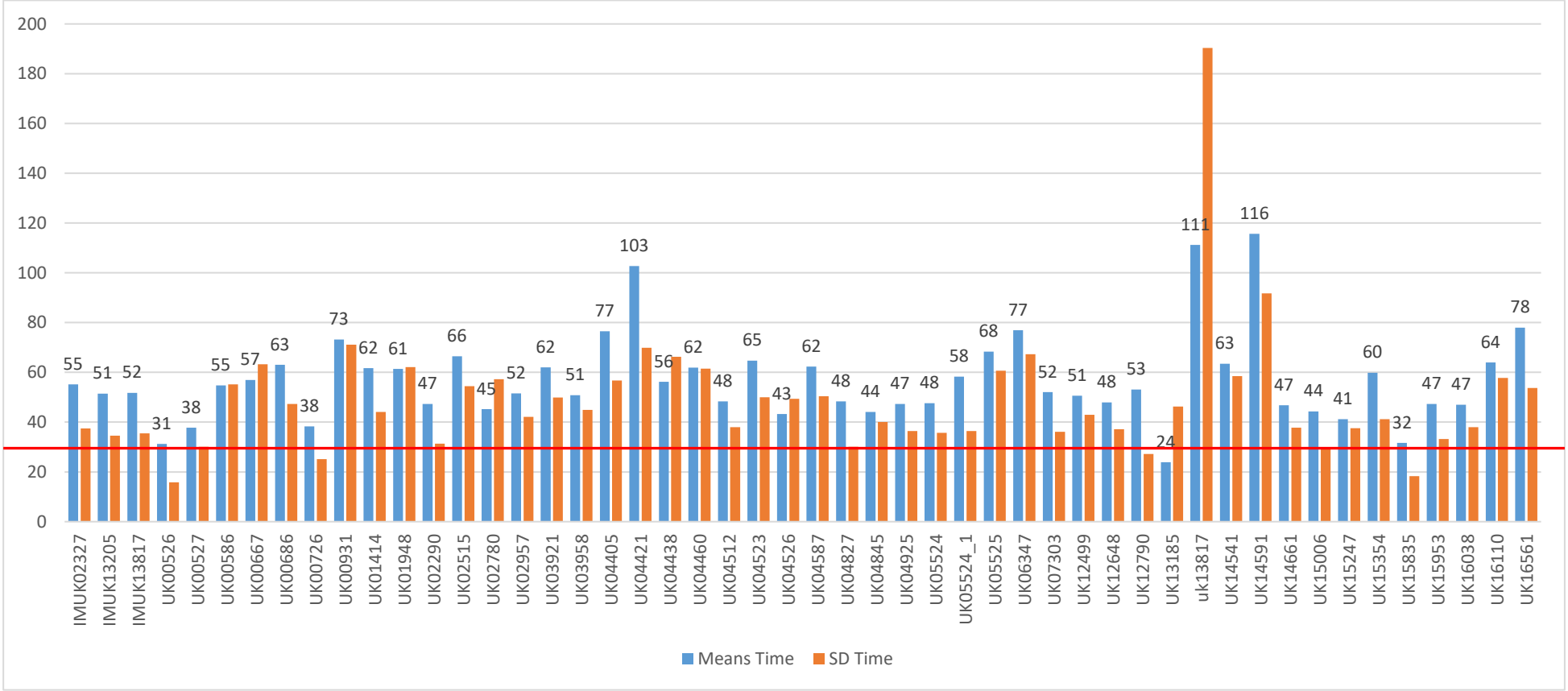
This figure shows the mean (and standard deviation) number of iterative steps used in the worse-than-dead element of the wheelchair example, by interviewer.

Figure 9. Use of WTD element in the TTO wheelchair example, by interviewer



This figure shows the total number of interviews by interviewer, and how often the worse-than-dead element of the wheelchair example was used. It is a protocol violation if the worse-than-dead element is not shown in the wheelchair example, because respondents may have worse than dead values, that remain hidden if the worse than dead task is not explained.

Figure 10. Time taken to complete a single TTO task, by interviewer



This figure shows the mean (and standard deviation) amount of time taken (in seconds) to complete each TTO task, by interviewer. This excludes the wheelchair example and practice TTO tasks. Completion of 10 TTO task in less than 30 seconds per state on average is considered a protocol violation. TTO valuation requires that the state is read aloud, that the interviewer confirms that the respondent is able to imagine the impact of the problems that are present on his life, and finally that a value is elicited. This process usually takes > 60 seconds per state to complete.

Interviewer effects

Figure 11. Mean values across all respondents and TTO tasks, by interviewer

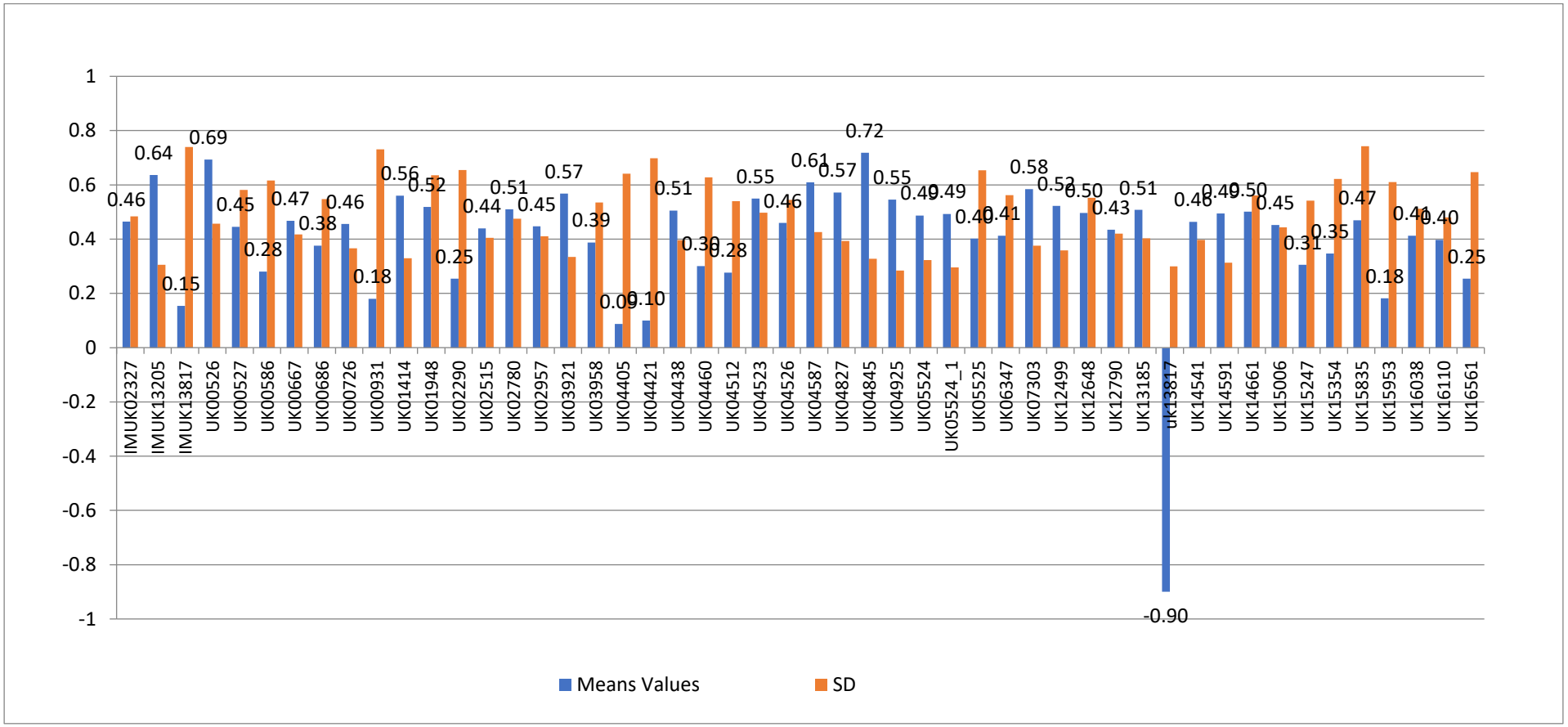


Figure 12. Non-traders, by interviewer

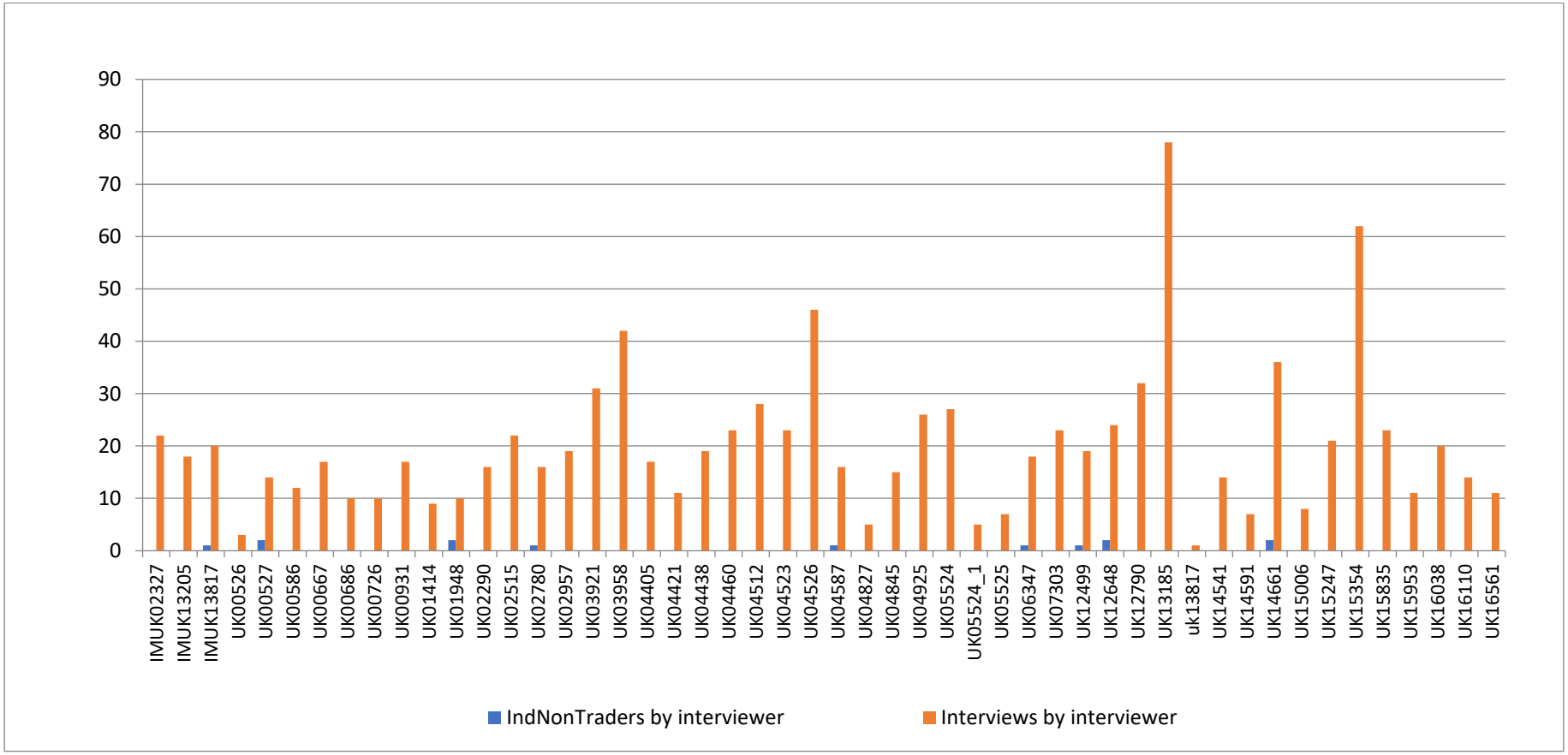


Figure 13. Number of 0 values, by interviewer

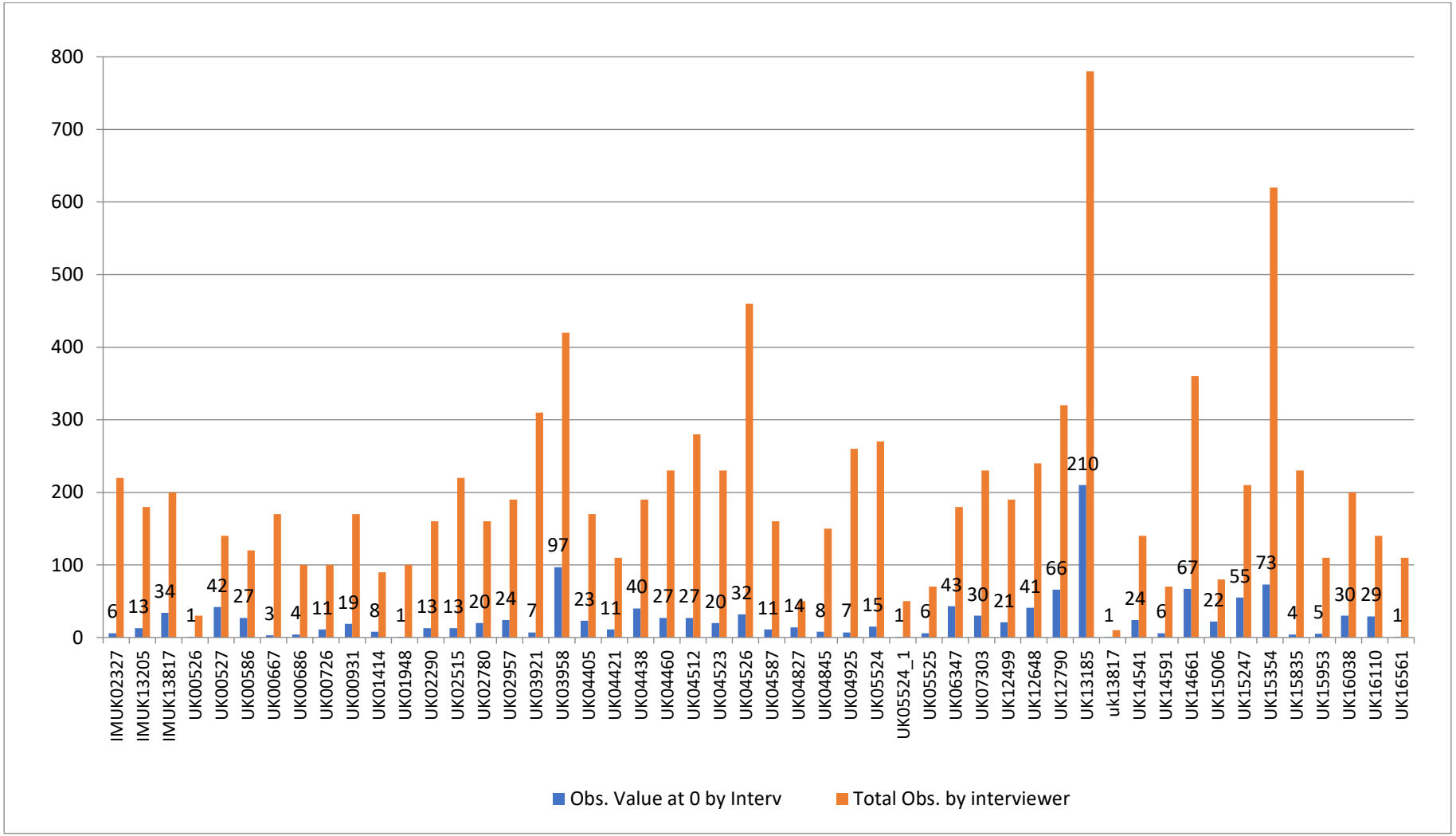


Figure 14. Number of negative values, by interviewer

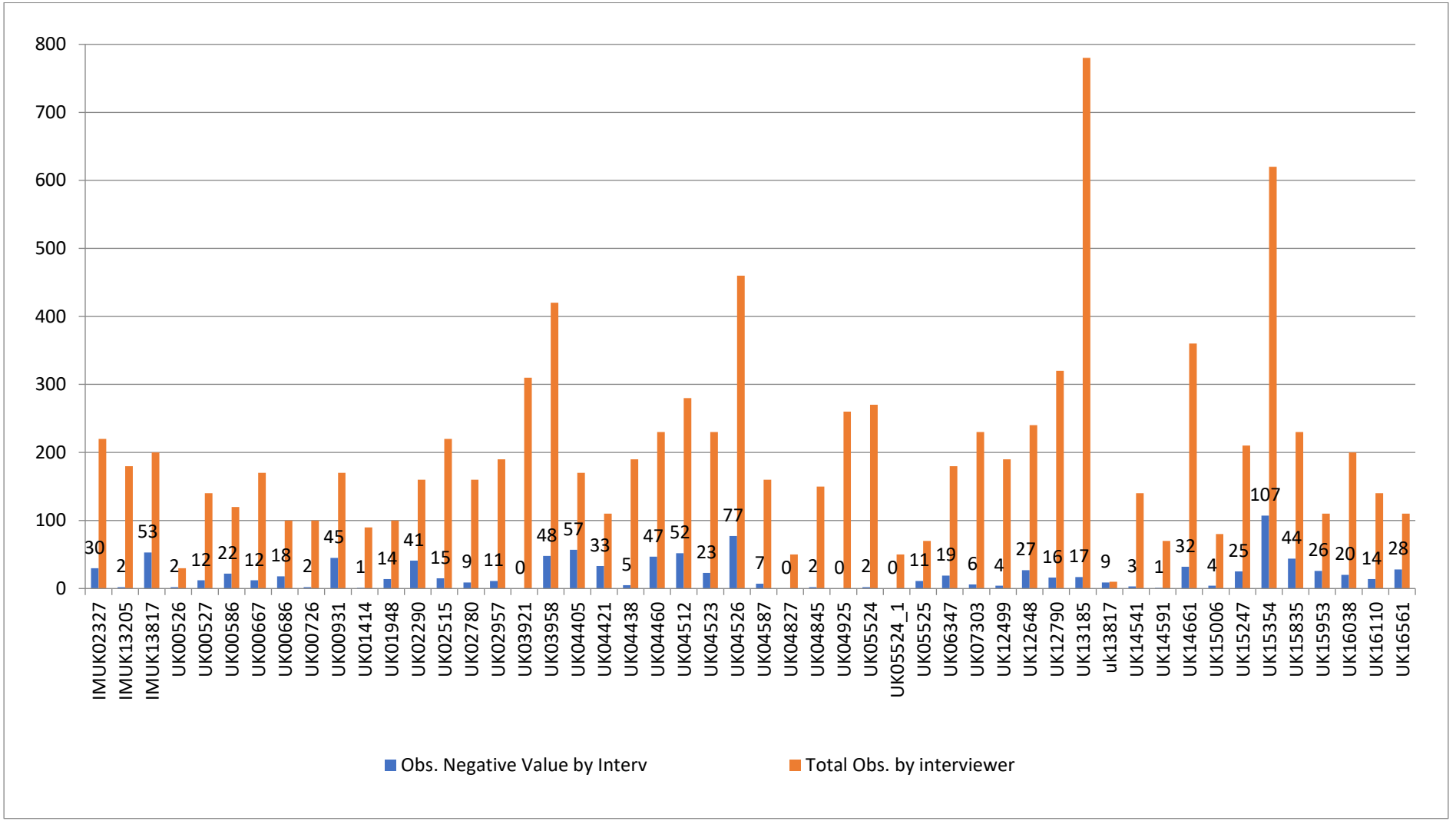
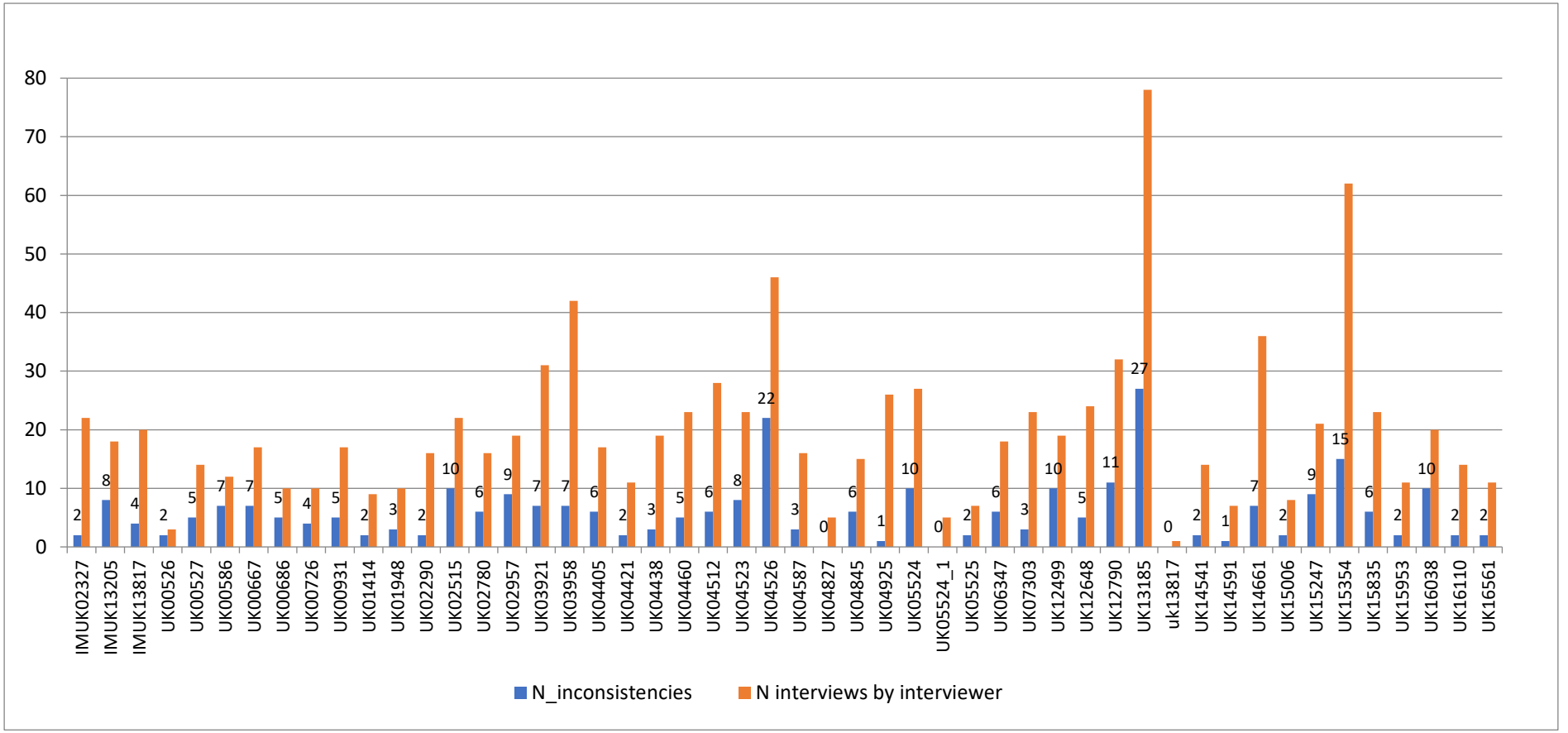


Figure 15. Inconsistencies by interviewer



This figure shows the number of respondents who valued state 55555 higher than at least one other state. State 55555 is the worst state in the EQ-5D-5L descriptive system. This state is included in all blocks and thus valued by all respondents. Out of the 10 states valued by an individual respondent, it is reasonable to expect that state 55555 has the lowest value. This will in practice not always be the case, for instance because random errors or learning effects interfere. However, if inconsistencies cluster in interviewer that is taken as a sign as low task engagement, which carries over to the respondent who might then complete the task half heartedly, and be more inclined to value states with low accuracy and much random error.

Figure 16. TTO value distribution for each interviewer

(one interviewer excluded who interviewed one person and that person valued all health states at -1)

Figure 16 shows the distribution of responses as collected by each interviewers.

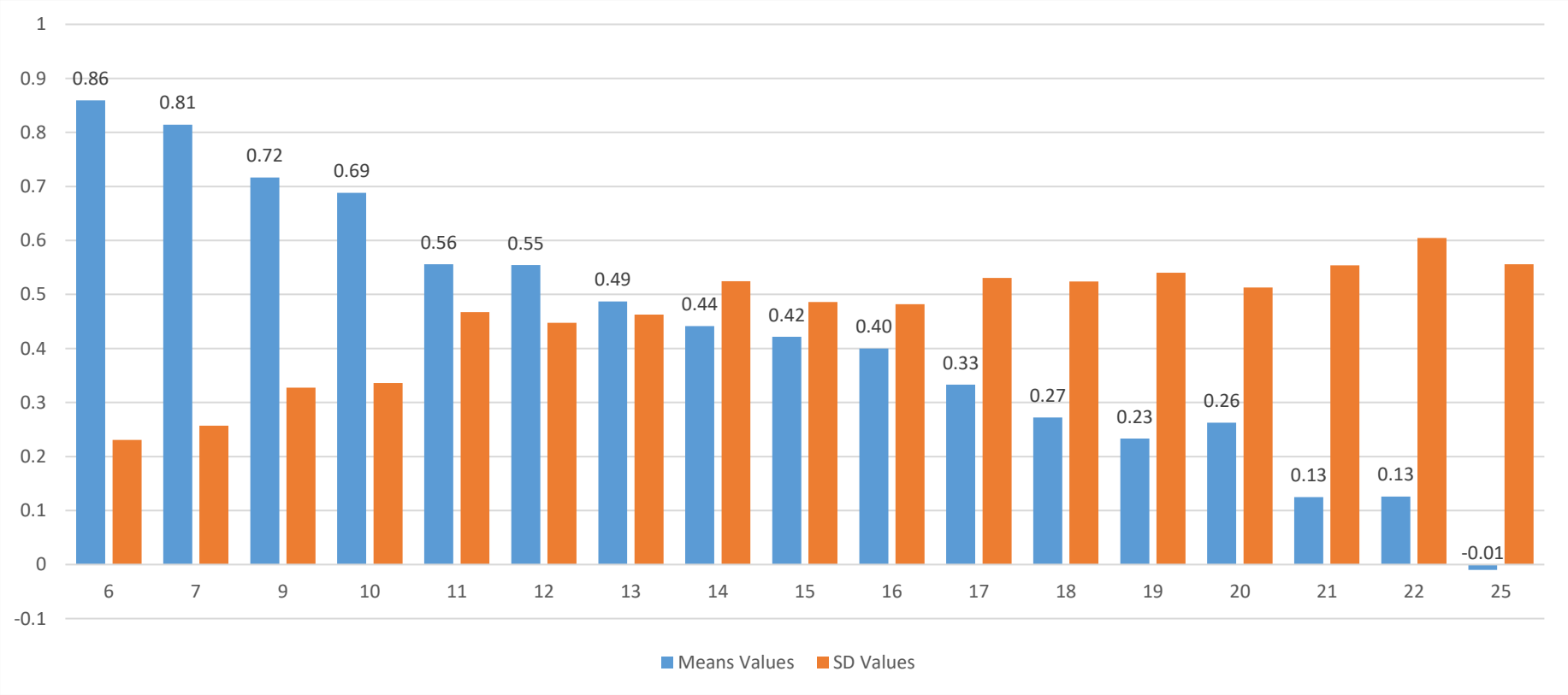
Each graph reports the data collected by one interviewer, across all respondents and health states. The X-axis identifies the range of attainable values and thus ranges from -1 to 1. The Y-axis reports the perecentage of responses at each of the attainable values. For every completed interview, 10 responses are included. For instance, the graphs of the two interviewers who completed the most interviews (78 & 62) include 780 and 620 observations respectively, and each observations represents the opinion of one respondent about one of the health states he had to value. On average, the interviewers conducted 20 interviews per person. The figure highlights the number of interviews done by interviewers who completed more than 40 interviews.

In this figure we observe

1. Few similarities across the individual graphs. This means that strong interviewer effects are present. It is likely that respondents would have given different answers if they had been interviewed by a different person.
2. Strong clustering effects in about half of the graphs. For instance, interviewer UK00527 (5th figure on the first row) completed 14 interviews and thus collected 140 responses. Observations cluster at 0 and 1 (together > 60%). This suggests that many health states have received the same value in spite of being obviously different. This also suggest an unlikely high level of agreement among respondents of how good or bad the health states are.
3. Strong clustering effects at 0 are often combined with few worse-than-dead values which may indicate that worse than dead values have been censored. This may for instance be illustrated by the difference between interviewers 13185 and 15953.
4. It is of little use to compare the distribution of responses of an individual interviewers to the distribution of all the total set of responses, a suggested benchmark, because data issues seem so widespread that the 'benchmark' is compromised.

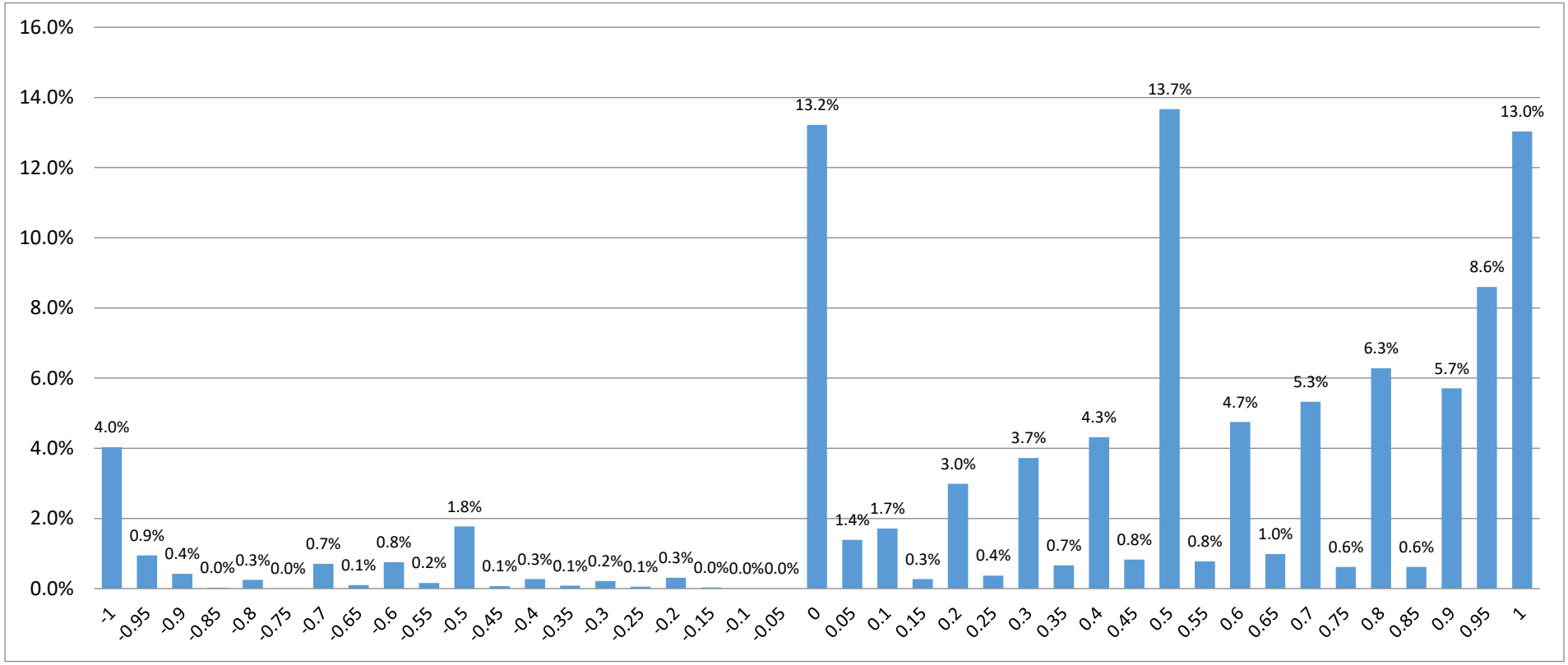
Face validity of aggregate data.

Figure 17. Mean TTO value, by level sum score



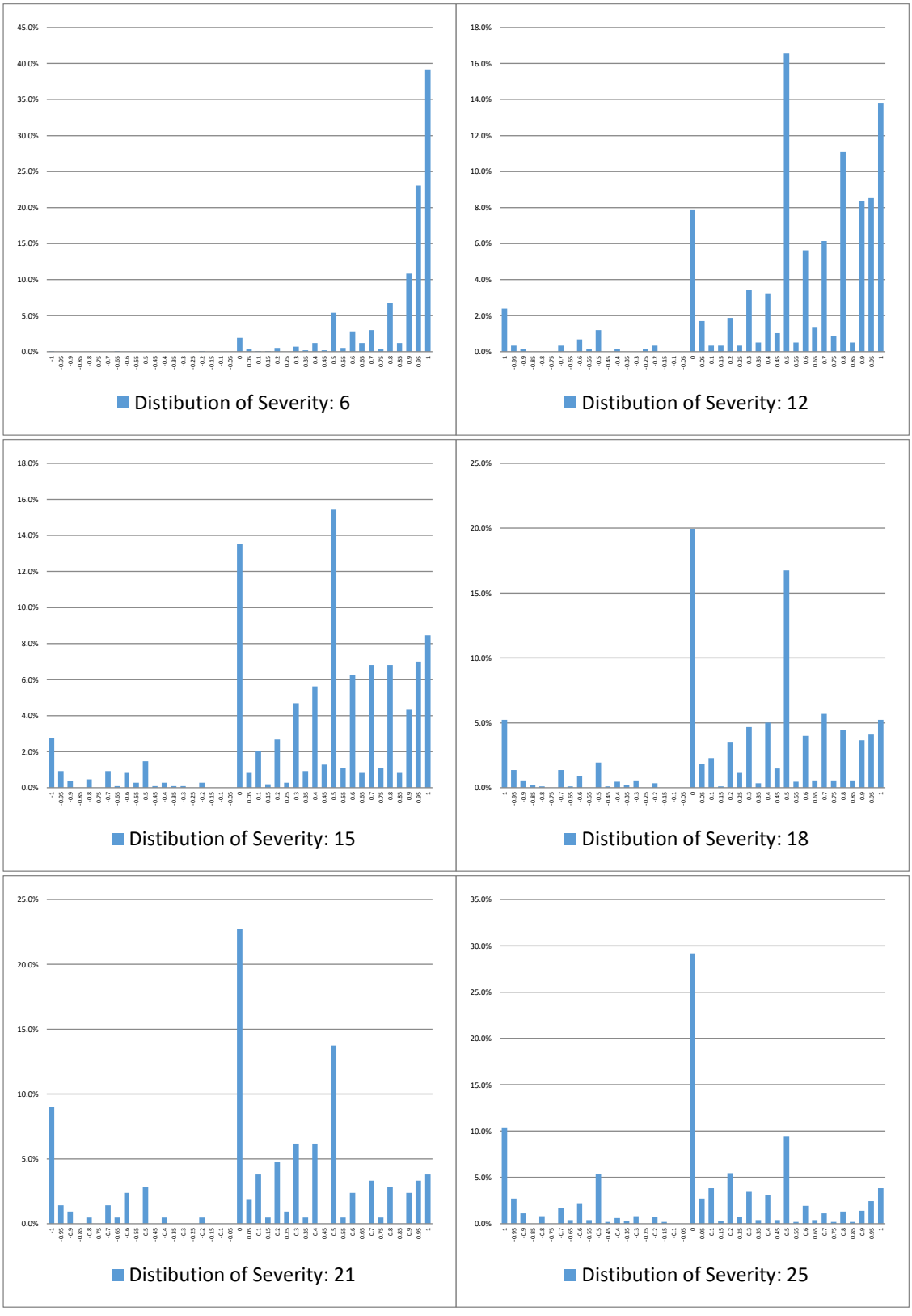
This figure shows the mean (and standard deviation) TTO value observed, by level sum score, across all interviewers. The level sum score is a proxy for severity and is calculated by summing the five dimension levels for each health state. We would expect health states with lower level sum scores (e.g. 21111: 2+1+1+1+1=6) to have higher mean values that those with higher level sum scores (e.g. 55555: 5+5+5+5+5=25). This excludes the wheelchair example and practice TTO tasks.

Figure 18. Overall TTO value distribution



This figure shows the 5L TTO value distribution for all health states. For example, the rightmost bar shows the proportion of observations of values greater than 0.95 and less than or equal to 1.0. This excludes the wheelchair example and practice TTO tasks.

Figure 19. TTO value distribution per different level sum score



These figures show the 5L TTO value distribution for health states with different level sum score (e.g. 6 will mean states like 21111). This excludes the wheelchair example and practice TTO tasks.

Table 3. DCE unusual responses

Interviewer	N	Time (min.)	IF AAAAAAA	IF BBBBBBB	IF ABABABA	IF BABABAB
IMUK02327	22	3.60	0	0	0	0
IMUK13205	18	3.77	0	0	0	0
IMUK13817	20	3.01	1	0	0	0
UK00526	3	2.60	0	0	0	0
UK00527	14	2.42	0	0	0	0
UK00586	12	3.57	0	0	0	0
UK00667	17	4.90	0	1	0	0
UK00686	10	3.36	0	0	0	0
UK00726	10	3.51	0	0	0	0
UK00931	17	3.84	0	0	0	0
UK01414	9	3.91	0	0	0	0
UK01948	10	4.63	0	1	0	0
UK02290	16	2.86	0	0	0	0
UK02515	22	6.01	0	0	0	0
UK02780	16	3.30	0	0	0	1
UK02957	19	4.13	0	0	0	0
UK03921	31	4.82	0	0	0	0
UK03958	42	3.35	1	0	1	0
UK04405	17	4.55	0	0	0	1
UK04421	11	5.75	0	0	0	0
UK04438	19	3.61	0	0	0	0
UK04460	23	3.98	1	0	0	0
UK04512	28	2.67	0	0	0	0
UK04523	23	4.17	0	0	0	0
UK04526	46	2.78	0	1	1	0
UK04587	16	3.41	0	0	0	0
UK04827	5	3.43	0	0	0	0
UK04845	15	3.43	0	0	0	0

Interviewer	N	Time (min.)	IF AAAAAAA	IF BBBBBBB	IF ABABABA	IF BABABAB
UK04925	26	4.51	0	0	0	0
UK05524	27	3.73	0	0	0	1
UK05524_1	5	5.30	0	0	1	0
UK05525	7	3.91	0	0	0	0
UK06347	18	4.22	0	0	0	0
UK07303	23	4.21	0	0	0	0
UK12499	19	4.40	0	1	0	0
UK12648	24	3.74	0	0	0	0
UK12790	32	3.02	0	0	1	0
UK13185	78	2.09	2	1	0	1
uk13817	1	3.94	0	0	0	0
UK14541	14	4.79	0	0	0	0
UK14591	7	7.01	0	0	0	0
UK14661	36	3.72	0	0	1	0
UK15006	8	3.25	0	0	0	0
UK15247	21	3.00	0	0	0	0
UK15354	62	3.85	0	0	0	0
UK15835	23	2.19	0	0	0	0
UK15953	11	3.81	0	0	0	0
UK16038	20	3.06	0	0	0	1
UK16110	14	4.50	0	0	0	0
UK16561	11	6.02	0	0	0	1

This table shows, by interviewer: the number of interviews completed (column 2); the mean amount of time taken (in minutes) to complete the 3 DC tasks (column 3); and the number of respondents who gave unusual sets of choices across all seven DCE tasks (columns 4-7). For example, if the respondent chose state A in all seven tasks, this is flagged in column 4.