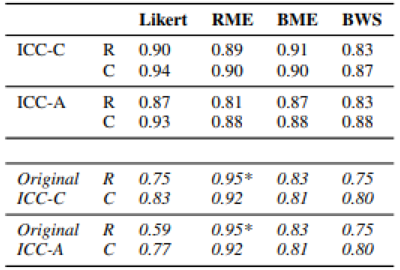
A REPRODUCTION STUDY OF METHODS FOR EVALUATING DIALOGUE SYSTEM OUTPUT: REPLICATING SANTHANAM AND SHAIKH (2019)

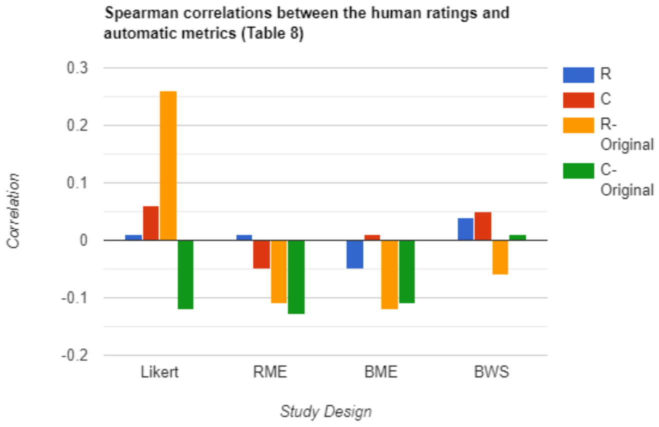
Anouck Braggaar, Frédéric Tomas, Peter Blomsma, Saar Hommes, Nadine Braun, Emiel van Miltenburg, Chris van der Lee, Martijn Goudbeek, Emiel Krahmer (Tilburg University)

Design	Participants	Average Time (minutes)
Likert Scale	42	58
Rank-Based Magnitude Estimation (RME)	40	54.7
Biased Magnitude Estimation (BME)	41	48.8
Best-Worst Scaling (BWS)	40	48.6

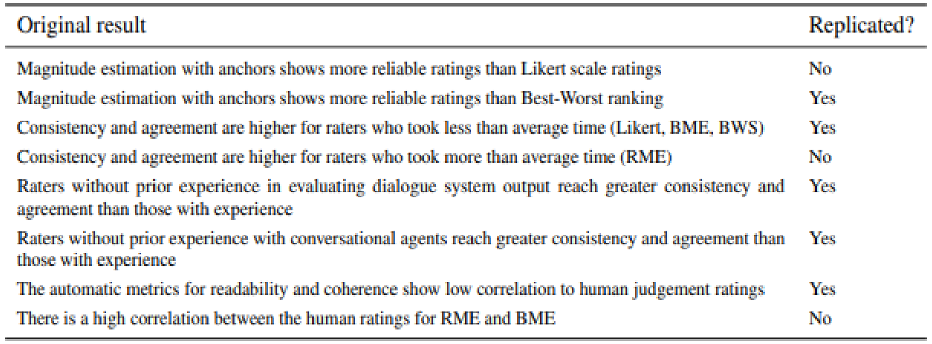


		Likert	RME	BME	BWS
ICC-C	R	0.90	0.89	0.91	0.83
	C	0.94	0.90	0.90	0.87
ICC-A	R	0.87	0.81	0.87	0.83
	C	0.93	0.88	0.88	0.88
Original ICC-C	*R*	*0.75*	*0.95**	*0.83*	*0.75*
	C	*0.83*	*0.92*	*0.81*	*0.80*
Original ICC-A	*R*	*0.59*	*0.95**	*0.83*	*0.75*
	C	*0.77*	*0.92*	*0.81*	*0.80*

Table 1: ICC scores for readability (R) and coherence (C) for each design. All are significant at $p < .001$. The original study scores are shown in italic with * showing the non-significant values.



Spearman correlations between the human ratings and automatic metrics (Table 8)



Original result	Replicated?
Magnitude estimation with anchors shows more reliable ratings than Likert scale ratings	No
Magnitude estimation with anchors shows more reliable ratings than Best-Worst ranking	Yes
Consistency and agreement are higher for raters who took less than average time (Likert, BME, BWS)	Yes
Consistency and agreement are higher for raters who took more than average time (RME)	No
Raters without prior experience in evaluating dialogue system output reach greater consistency and agreement than those with experience	Yes
Raters without prior experience with conversational agents reach greater consistency and agreement than those with experience	Yes
The automatic metrics for readability and coherence show low correlation to human judgement ratings	Yes
There is a high correlation between the human ratings for RME and BME	No

Table 11: Results evaluated for replicability in this paper.

Take-aways

- Results generally replicate (3/8 of all results)
- Check reliability of participants
- Share and document all code
- Standardize surveys