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** and **Emiel Krahmer**
Tilburg University
✉ A.R.Y.Braggaar@tilburguniversity.edu

Abstract

In this paper, we describe our reproduction effort of the paper: *Towards Best Experiment Design for Evaluating Dialogue System Output* by Santhanam and Shaikh (2019) for the 2022 ReproGen shared task. We aim to produce the same results, using different human evaluators, and a different implementation of the automatic metrics used in the original paper. Although overall the study posed some challenges to reproduce (e.g. difficulties with reproduction of automatic metrics and statistics), in the end we did find that the results generally replicate the findings of Santhanam and Shaikh (2019) and seem to follow similar trends.

1 Introduction

Currently, a lot of attention is given to the reproducibility of NLP research. In this paper, we report our contributions to the 2022 ReproGen shared task (Belz et al., 2020).[1] We aim at an exact reproduction of the work by Santhanam and Shaikh (2019) on experiment design for evaluating dialogue system output. No other reproductions of this paper have been published presently. We will first give a brief summary of the paper we aimed to reproduce (§2), and explain how we replicated this research as closely to the original as possible (§3). Next, we will discuss our results and examine how these relate to the original study (§4). Lastly, we will discuss some difficulties we faced during our reproduction efforts (§5). All of our code and data can be found on GitHub.[2]

2 Summary of the original study

The study by Santhanam and Shaikh (2019) focuses on the design of the human evaluation task of evaluating dialogue system output. The purpose of the task is to see which task design yields the most consistent and highest-quality responses. The original study compared Likert scale judgments, Rank-Based Magnitude Estimation (RME), Biased Magnitude Estimation (BME) and Best-Worst Scaling (BWS) using the two metrics of readability and coherence.

Participants. The authors examined four different experimental conditions with 40 participants and 50 items each, yielding a total number of 160 participants.

Task. For each trial, participants were provided with a conversational context consisting of two turns. For each context, participants were asked to either rank or rate four different responses. Three of these responses were generated by three models trained on the Reddit conversation corpus (Dziri et al., 2019a). The other response was human-generated (i.e. the ground truth). In case of the Likert scale, people rate a generated response on a 6-point scale (1 being lowest and 6 highest). For both RME and BME (magnitude estimation) they rate the responses with respect to a given standard value (Bard et al., 1996). In the case of RME this value is always 100 while the value for BME is set by the automatic metrics of Santhanam and Shaikh (2019). In the last experiment design, BWS, participants have to rank the responses from best to worst.

Reported values. Santhanam and Shaikh (2019) report on inter-rater consistency and agreement (intraclass correlations) and also examine if prior experience of rating dialogue system output or engaging with a conversational agent is of any influence. Lastly, Spearman correlations are reported between the human ratings and automatic metrics and between the ratings of readability and coherence on the four designs.

Results. Overall, Santhanam and Shaikh (2019) find that the Likert scale performs worst on intraclass correlation, show that participants without prior experience are more consistent in their ratings,

[1] https://reprogen.github.io/
[2] https://github.com/Anouck96/ReproGen22

Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges, pages 86 - 93
July 17-22, 2022

and report low correlation between the automatic metrics and the human ratings.

3 Study design

3.1 Surveys

Our surveys were made in the online survey platform Qualtrics.[3] We tried to follow the survey design of the original study as closely as possible. Unfortunately, in some cases, the layout was not exactly replicable. An example can be found in the best-worst scaling condition. Qualtrics neither provides the same drag-and-drop ranking question types as used in the original survey nor does it track if an item is ranked or not. We also found that the four original surveys are not completely the same in terms of conversation items and possible replies, e.g., in the Likert survey, the item "Person A: first time watching f1!" occurs twice, while in the best-worst survey, the item only occurs once. There were also some minor layout/style issues that we noticed. For example, some questions in the best-worst survey contained "readability and coherence" in bold while others did not.

3.2 Participants

Participants were recruited using Prolific[4], a crowdsourcing recruitement platform. Only participants with English as their first language could take part in the study. The participants also were not allowed to participate in more than one of the surveys, beyond this they were also not allowed to participate in the same survey twice. We followed the minimum payment of £6.00 per hour resulting in rewards for the participants of £5.43 (Likert), £4.74 (RME), £4.64 (best-worst) and £4.88 (BME), as prolific uses the median time for payments. As in the original study, we aimed for 40 participants per survey. We started with the Likert-scale survey on Prolific. We set a time based on the mean times reported in Santhanam and Shaikh (2019) (as required by prolific), but we soon discovered that participants tended to take substantially more time, with a median completion time of about 53 minutes. In the case of the Likert scale survey, we even had participants who timed out[5] but were able to finish the survey.[6] These were kept in the dataset. This is why the Likert survey contains 42 participants. For the BME survey, we have 41 participants. In this case we have two submissions with the same Prolific ID but with different answers. As they do have different answers we have decided to keep both submissions. For the other two surveys, we have 40 submissions.

3.3 Intraclass correlation coefficient (ICC)

Intraclass correlation is used to calculate the reliability of raters (Bartko, 1966). In this study we report values for both agreement and consistency. The values can range between 0 and 1 (closer to 1 means stronger reliability) (Koo and Li, 2016). The ICC was calculated using R (R Core Team, 2017) and the irr package (Gamer et al., 2019).

3.4 Automatic metrics

To fully replicate the original results, we recalculated the automatic metrics used by Santhanam and Shaikh (2019). Since the repository did not provide any code to generate the scores, we first contacted the authors to obtain the exact code that was used for the original paper. However, at the time of writing, the repository did not provide any code to generate the scores. Hence we needed to write our own code to process the data and generate the scores ourselves.

The original paper did not specify what library they used to compute readability. Thus, we explored different options to generate the exact same readability scores.[7] In the end, we did not find an exact match. We decided to calculate the Flesch Reading Ease using the `textacy` Python package.[8] For coherence, Santhanam and Shaikh noted that they used the method proposed by Dziri et al. (2019b). We used their repository to calculate the semantic similarity.[9]

4 Results

In this section, we follow the original study's approach in the data analysis and its structure in the

[3] See: https://www.qualtrics.com. PDF files with the surveys as they were used in Qualtrics can be found on GitHub. We do not know which platform was used for the questions in the original study, although it seems likely the authors used the Mechanical Turk platform itself.

[4] https://www.prolific.co/

[5] To ensure fair payment prolific sets a maximum time based on the set time for the study.

[6] Two participants timed out which meant they were automatically replaced by Prolific, however their completed surveys were collected in Qualtrics.

[7] We looked into textstat, py-readability-metrics, and Microsoft Word, which all generate different readability scores.

[8] https://textacy.readthedocs.io/en/latest/

[9] https://github.com/nouhadziri/DialogEntailment

organization of our results.

4.1 Experiment design and reliability of human ratings

Intraclass correlation coefficient (ICC) scores on consistency (ICC-C) and agreement (ICC-A) for the four experiment tasks can be found in Table 1. Unlike the findings reported in Santhanam and Shaikh (2019), Magnitude Estimation with anchors (RME or BME) does not show more reliable ratings than Likert scale ratings, but it does show more reliable ratings than Best-Worst ranking (BWS). Likert scale ratings result in substantially higher ICC scores in our replication. In fact, the Likert scale condition leads to the most reliable ratings, while Best-Worst ranking (BWS) represents the least reliable ratings in our results. With the exception of RME, all experimental designs show higher ICC scores in our study.

		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.

4.2 Time and reliability of the rankings

As mentioned in section 3.2, participants in our replication study took a median completion time for the Likert-survey of about 53 minutes, which substantially exceeds the averages reported in the original study (see 5.1 for a more elaborate discussion on experiment times). Table 2 contains the ICC scores for raters who spent more than average time on the task, and Table 3 contains the ICC scores for raters who spent less than average time.

We replicate the finding of Santhanam and Shaikh (2019) that consistency and agreement are higher for raters who took less than average time to complete the task, but in all survey conditions, including RME. The RME survey showed the opposite direction in the original study. Additionally, we

		Likert (n=9)	**RME** (n=16)	**BME** (n=19)	**BWS** (n=17)
ICC-C	R	0.68	0.64	0.82	0.66
	C	0.74	0.70	0.79	0.71
ICC-A	R	0.60	0.47	0.75	0.67
	C	0.72	0.66	0.75	0.71
		n=15	*n=16*	*n=15*	*n=16*
Original ICC-C	*R*	*0.58*	*0.93*	*0.51*	*0.62*
	C	*0.74*	*0.85*	*0.55*	*0.64*
Original ICC-A	*R*	*0.52*	*0.93*	*0.51*	*0.62*
	C	*0.69*	*0.86*	*0.56*	*0.64*

Table 2: ICC scores for readability (R) and coherence (C) where participants spend **above** average time. All are significant at $p < .001$. Original study scores are in italics.

		Likert (n=33)	**RME** (n=24)	**BME** (n=22)	**BWS** (n=23)
ICC-C	R	0.88	0.88	0.86	0.74
	C	0.93	0.89	0.83	0.83
ICC-A	R	0.83	0.79	0.80	0.74
	C	0.92	0.86	0.80	0.83
		n=25	*n=24*	*n=25*	*n=24*
Original ICC-C	*R*	*0.61*	*0.88*	*0.81*	*0.65*
	C	*0.66*	*0.85*	*0.75*	*0.76*
Original ICC-A	*R*	*0.36*	*0.88*	*0.81*	*0.66*
	C	*0.55*	*0.85*	*0.75*	*0.76*

Table 3: ICC scores for readability (R) and coherence (C) where participants spend **below** average time. All are significant at $p < .001$. The original study scores are in italic.

observe different patterns: the RME condition led to the highest reliability in the original study, both for raters taking above and below average time. In our study, RME actually leads to the lowest reliability for raters taking above average time (the highest being BME), and Likert scale ratings lead to the highest reliability for raters taking below average time (lowest in the original study).

4.3 Prior experience with dialogue system output or conversational agents and reliability of rankings

Tables 4 and 5 show the reliability scores of ratings from participants based on their prior experience with dialogue-system output evaluation. We replicate the findings reported in the original study: ratings from participants without prior experience with evaluating dialogue system output reach better

reliability than ratings from participants with such prior experience. We also replicate that no prior experience with conversational agents benefits the consistency and reliability of participants' ratings (Tables 6 & 7).

		Likert (n=10)	**RME** (n=5)	**BME** (n=8)	**BWS** (n=5)
ICC-C	R	0.74	0.75	0.65	0.28*
	C	0.71	0.64	0.60	0.42
ICC-A	R	0.64	0.72	0.55	0.28*
	C	0.67	0.61	0.52	0.42
		n=15	*n=7*	*n=18*	*n=13*
Original ICC-C	*R*	*0.45*	*0.37*	*0.51*	*0.54*
	C	*0.38*	*0.48*	*0.55*	*0.63*
Original ICC-A	*R*	*0.35*	*0.38*	*0.52*	*0.55*
	C	*0.32*	*0.49*	*0.55*	*0.63*

Table 4: ICC scores for readability (R) and coherence (C) when participants **have** prior experience evaluating dialogue system output. All are significant at $p < .001$, except those indicated with *. Original study scores in italic.

		Likert (n=32)	**RME** (n=35)	**BME** (n=33)	**BWS** (n=35)
ICC-C	R	0.88	0.87	0.89	0.81
	C	0.93	0.89	0.88	0.86
ICC-A	R	0.84	0.77	0.84	0.81
	C	0.92	0.86	0.85	0.86
		n=25	*n=33*	*n=22*	*n=27*
Original ICC-C	*R*	*0.71*	*0.95**	*0.83*	*0.70*
	C	*0.82*	*0.92*	*0.76*	*0.72*
Original ICC-A	*R*	*0.50*	*0.95**	*0.83*	*0.70*
	C	*0.75*	*0.92*	*0.77*	*0.72*

Table 5: ICC scores for readability (R) and coherence (C) when participants **do not have** prior experience evaluating dialogue system output. All are significant at $p < .001$. The original study scores are shown in italics with * showing the non-significant values.

4.4 Correlation of automated calculation of readability and coherence with human ratings

Santhanam and Shaikh (2019) found low correlations between the automatic metrics and human judgements, ranging from -0.12 to 0.26. We find even lower correlations between readability and coherence scores calculated with automated methods and human ratings (see Table 8).

		Likert (n=16)	**RME** (n=15)	**BME** (n=16)	**BWS** (n=13)
ICC-C	R	0.80	0.62	0.78	0.52
	C	0.86	0.77	0.79	0.57
ICC-A	R	0.72	0.46	0.66	0.52
	C	0.83	0.73	0.74	0.58
		n=18	*n=11*	*n=23*	*n=18*
Original ICC-C	*R*	*0.46*	*0.69*	*0.60*	*0.57*
	C	*0.44*	*0.65*	*0.62*	*0.67*
Original ICC-A	*R*	*0.37*	*0.69*	*0.61*	*0.57*
	C	*0.38*	*0.65*	*0.62*	*0.67*

Table 6: ICC scores for readability (R) and coherence (C) when participants **have** prior experience engaging with conversational agents. All are significant at $p < .001$. Original study scores in italic.

		Likert (n=26)	**RME** (n=25)	**BME** (n=25)	**BWS** (n=27)
ICC-C	R	0.85	0.89	0.87	0.78
	C	0.91	0.86	0.84	0.85
ICC-A	R	0.80	0.80	0.83	0.78
	C	0.89	0.83	0.81	0.85
		n=22	*n=29*	*n=17*	*n=22*
Original ICC-C	*R*	*0.70*	*0.95**	*0.84*	*0.67*
	C	*0.82*	*0.91*	*0.76*	*0.68*
Original ICC-A	*R*	*0.48*	*0.95**	*0.84*	*0.67*
	C	*0.75*	*0.91*	*0.76*	*0.68*

Table 7: ICC scores for readability (R) and coherence (C) when participants **do not have** prior experience engaging with conversational agents. All are significant at $p < .001$. Original study scores in italics with * showing non-significant values.

	Likert	**RME**	**BME**	**BWS**
Readability	0.01	0.01	-0.05	0.04
Coherence	0.06	-0.05	0.01	0.05
Original scores				
Readability	*0.26*	*-0.11*	*-0.12*	*-0.06*
Coherence	*-0.12*	*-0.13*	*-0.11*	*0.01*

Table 8: Spearman correlation between the ratings obtained from the automated metrics to human ratings using raw scores. Original study scores in italic.

4.5 Correlation of readability and coherence by experiment condition

We do not replicate the high correlations between the human ratings of readability and coherence obtained through RME and BME (see Spearman correlations in Table 9). For Likert, RME, and

BME, correlations are weak, while similar to the original paper, we find a moderate correlation for human ratings obtained through BWS.

	Likert	**RME**	**BME**	**BWS**
		Readability		
Coherence	0.13*	0.06	0.24**	0.48***
Original				
	0.1	*0.79***	*0.77***	*0.5***

Table 9: Spearman correlation between the ratings obtained for readability and coherence for each human evaluation method, *$p \leq .05$, **$p \leq .01$, ***$p \leq .001$. Original scores in italic.

		Likert	**RME**	**BME**	**BWS**
Readability	Mean	0.64	0.39	0.47	0.61
	Mode	0.22	0.66	0.36	0.49
Coherence	Mean	0.82	0.77	0.61	0.72
	Mode	0.39	0.57	0.31	0.48

Table 10: Correlations between the original results and the reproduction study results. The correlations were calculated on the average and modal score for each sentence, respectively. All $p \leq .001$.

5 Discussion

The purpose of this study was to gain insights that can aid the Natural Language Generation (NLG) community to increase reproducibility of papers, specifically papers regarding human and automatic evaluation of NLG results. We reproduced the work from Santhanam and Shaikh (2019) including the experiments and the analyses. The results from Santhanam and Shaikh (2019) and our reproduction are equivalent, or at least in the same order of magnitude. Table 10 displays the correlation between their results and ours regarding readability and coherence across all four measures, indicating that, mostly, their measures and ours seem to correlate quite importantly. Additionally, Table 11 discloses an overview of the results. Below we discuss observations and insights gained during this reproduction exercise.

5.1 Participants

As mentioned before in Section 3.2, we used the average time that participants took to calculate our budget on Prolific. As we found out soon with running the first survey, participants took way longer than the estimated average. Our participants took an average of approximately 58 minutes for the Likert-scale survey (SD=24.47), 54.7 minutes for RME (SD=23.39), 48.8 for BME (SD=18.68) and 48.6 for best-worst ranking (SD=22.31). Santhanam and Shaikh (2019) report averages respectively of 33, 42.8, 43 and 32.5 minutes. As can be seen from our standard deviations, the amount of time also varied greatly across participants. Our participants especially seem to take much longer for the Likert and best-worst surveys. We are not sure why the difference is this large. With an online survey where there is no supervision, it is possible that participants get distracted or take breaks during the experiment. Therefore, averages could be lower in a lab-setting where participants are only focused on the task. Other options would be that we just recruited slower participants, or that the Qualtrics survey design makes it more difficult to answer quickly.

5.2 Response quality

Output quality for any annotation task depends on three factors: clarity of the task, ambiguity of the items, and the reliability of the annotators (Aroyo and Welty, 2014). Here we focus on the latter. Not all participants are equally reliable in their responses. If we assume that there is one true ranking or quality score for each dimension,[10] one reasonable way to approximate this true value is to take the average of all responses. We used this intuition to measure the reliability of each participant's scores by comparing their scores to the average scores of all other participants for each item. Figure 1 shows the results for the different metrics.

We observe that there is a fair (0.33) to moderate (0.64) correlation between participants' reliability scores for the relevance and coherence scales. This means that participants who agreed more with other participants on one dimension, also tended to agree more with participants on the other dimension.

We also observe that for each metric, there is a nonzero amount of participants who obtained a Spearman correlation of zero or less with the other participants. We did not exclude any participants from our analysis, to stay true to Santhanam and Shaikh's original report, but depending on the context, one may want to exclude participants who fall below a certain threshold, to obtain a more re-

[10]This may in fact depend on the ambiguity of the item or the perspective of the annotator (Basile et al., 2021), but for this task we believe that we can make this simplifying assumption.

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.

liable estimate of output quality (again assuming that there is a single, 'True' quality score that we aim to estimate).

Finally, the distribution of the participant reliability scores seems to differ between metrics. For example, while the participants' reliability scores for the Likert scale seems to cluster together in the top right corner, the RR scores seem to be spread out more.

5.3 Automatic metrics

Another issue that we struggled with was the reproduction of the automatic metrics. While we followed the original paper's descriptions, the calculation of these automatic metrics was not completely clear and resulted in large differences between results. As we had some values from the original study (in their BME-survey), we could compare our metrics to theirs, but we never figured out how to consistently extract the same results. Next to the calculation of the automatic metrics themselves, we were also unsure how the rankings were derived from these metrics. This was not explicitly mentioned in the paper or the supplementary material. Finally, we discovered that they seemed divided into a 25/25/25/25 split. For future work we would suggest to use the code of the original paper for the reproduction of the automatic metrics.

5.4 Standardisation of surveys

To upload the surveys in our survey platform, we had to redesign and retype all four surveys from the supplied PDF files. This task took about four hours per survey. Such a retyping task is a barrier to perform a reproduction, and increases the risk of introducing typos into the surveys. Therefore, we recommend researchers to not only share the PDF files of their original survey, but also other available formats (in case of Qualtrics, the QSF format), such that the retyping task can be prevented.[11]

5.5 Statistical analyses

The statistical analyses were tedious as, despite the sharing of the data and the RMarkdown files, some transformations had been operated on the raw data (i.e., data conversion from raw scores to what we assumed to be ranked scores for BME and RME measures). We could not replicate these transformations despite multiple attempts to contact the authors. We thus ran our statistical analyses based on our own raw data and found the above-mentioned results.

5.6 Study-specific remarks

Santhanam and Shaikh (2019) show that the same content evaluated by four different types of evaluation tasks lead to four different outcomes. The outcomes within each task have a high correlation (high ICC scores). However, the correlation between the outcomes across the evaluation tasks is low. This is possibly because Likert allows for more degrees of freedom in answering a question. A question contains one utterance and four different replies that have to be rated on a 6-point Likert scale. Such question can be answered in $6^4 = 1296$ different ways. In comparison, the best-worst scaling evaluation task allows only $4! = 24$ different ways to answer the same question. Therefore, one would expect a higher ICC for the outcomes of best-worst scaling than those of the Likert evaluation

[11] As far as we know, different survey platforms (SurveyMonkey, Qualtrics, Google Forms, Alchemer) do not have a standard survey file format implemented yet, so some amount of conversion may still be necessary.

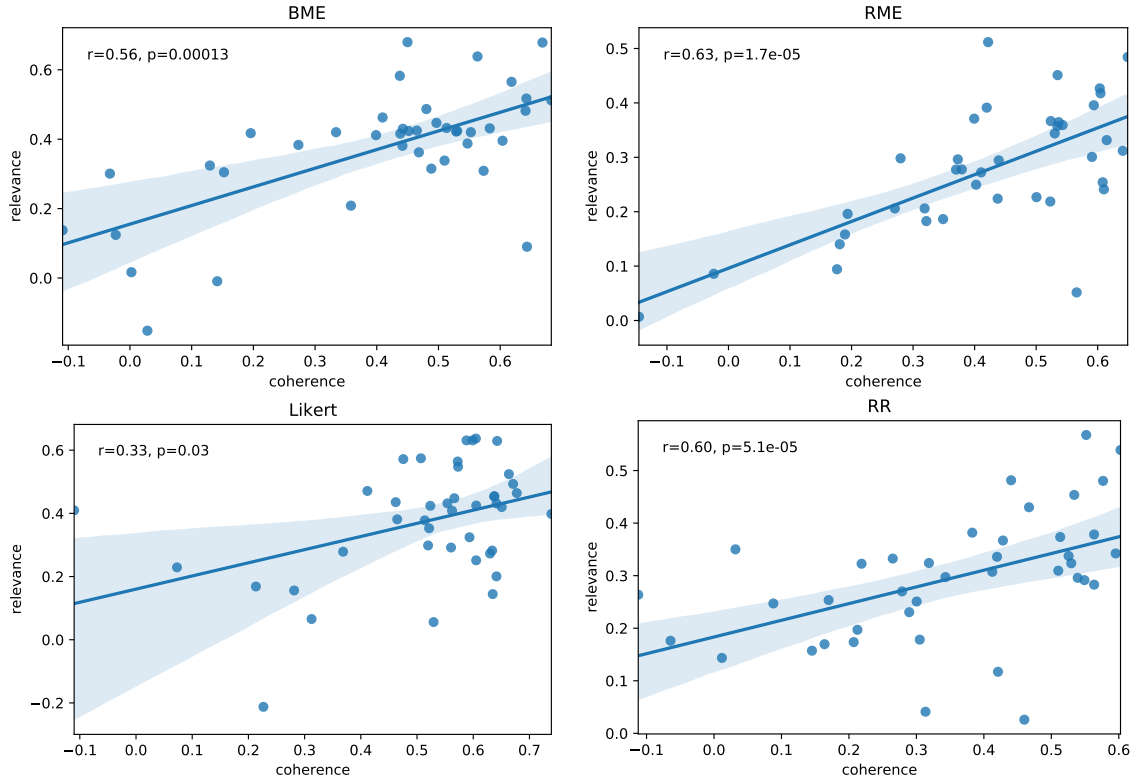

Figure 1: Scatterplot for the correlation between each participant's scores and the average of the other participants' scores. Each point represents one participant, and the axes correspond to the quality dimensions. In other words: these plots show a second-order correlation, measuring whether the reliability of participants (measured as correlations between each participant's scores and the average of other participants' scores) correlates between the two quality dimensions.

task. Furthermore, if we assume that some questions have only low quality replies, then a participant can express that within the Likert evaluation task, but in the best worst scaling task, the participant has to choose a best reply (even if such reply does not exist). The RME and BME evaluation tasks allow an average score. However, the Likert evaluation task is on a 6-point scale, so the participant is forced to evaluate each reply as slightly bad or slightly good. This could influence the correlations between the outcomes of the Likert evalation task on the one hand, and the RME and BME evalation tasks, on the other hand.

6 Conclusion

In this paper, we have tried to reproduce the work of Santhanam and Shaikh (2019). Our results generally replicate the findings of Santhanam and Shaikh (2019) and seem to follow similar trends. As discussed in Section 5, we did run into some difficulties throughout the reproduction process. We hope that our observations are instructive for future researchers in making their work fully reproducible.

Acknowledgements

Funding for paying the participants is supported by the NWO-funded project 628.011.030 Data2Person. Anouck Braggaar's contribution was supported by the NWO Smooth Operators project (KIVI.2019.009).

This research was approved by the Research Ethics and Data Management Committee of the Tilburg School of Humanities and Digital sciences (Tilburg University). Reference number: REDC 2019.40abc (amendment April 5, 2022).

References

Lora Aroyo and Chris Welty. 2014. The three sides of crowdtruth. *Human Computation*, 1(1).

Ellen Gurman Bard, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68.

John J Bartko. 1966. The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, 19(1):3–11.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the the 18th conference of the Italian Chapter of AIS (Association for Information Systems)*. Available through http://www.itais.org/itais2021-proceedings/pdf/21.pdf or https://arxiv.org/abs/2109.04270.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2020. ReproGen: Proposal for a shared task on reproducibility of human evaluations in NLG. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 232–236, Dublin, Ireland. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019a. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31, Florence, Italy. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019b. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3806–3812, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthias Gamer, Jim Lemon, and Ian Fellows Puspendra Singh <puspendra.pusp22@gmail.com>. 2019. *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1.

Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94, Tokyo, Japan. Association for Computational Linguistics.