How reproducible is best-worst scaling for human evaluation? A reproduction of 'Data-to-text Generation with Macro Planning'

Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Debby Damen, Martijn Goudbeek, Chris van der Lee, Frédéric Tomas, Emiel Krahmer
Department of Communication and Cognition
TiCC - Tilburg center for Cognition and Communication
Tilburg University
C.W.J.vanMiltenburg@tilburguniversity.edu

Abstract

This paper is part of the larger ReproHum project, where different teams of researchers aim to reproduce published experiments from the NLP literature. Specifically, ReproHum focuses on the reproducibility of human evaluation studies, where participants indicate the quality of different outputs of Natural Language Generation (NLG) systems. This is necessary because without reproduction studies, we do not know how reliable earlier results are. This paper aims to reproduce the second human evaluation study of Puduppully and Lapata (2021), while another lab is attempting to do the same. This experiment uses best-worst scaling to determine the relative performance of different NLG systems. We found that the *worst* performing system in the original study is now in fact the *best* performing system across the board. This means that we cannot fully reproduce the original results. We also carry out alternative analyses of the data, and discuss how our results may be combined with the other reproduction study that is carried out in parallel with this paper.

1 Introduction

Although human evaluation remains the gold standard for determining the quality of a text, little is known about the reproducibility of evaluation methods that are used to determine the quality of texts generated by Natural Language Generation (NLG) systems (Belz et al., 2020). To be sure, there are many different ways to assess NLG output. As van der Lee et al. (2019) and Howcroft et al. (2020) showed: different authors tend to focus on different quality dimensions (e.g. *grammaticality, coherence, conciseness*) and they also differ in the way they elicit quality judgments (e.g. using Likert scales or ranking tasks). If we want to understand how reliable these methods are, we need to carry out multiple reproduction studies to establish the amount of variance we may expect between different studies (Belz, 2022).

1.1 The ReproHum project

Establishing the reproducibility of human evaluation metrics is a relatively slow and incremental process, as it takes a great deal of time and resources to exactly reproduce even a single study. However, with a collective effort, we are currently making headway to achieve this goal. As part of the ReproHum project (Belz et al., 2023), this paper aims to reproduce an experiment from a published NLG study, while another lab (identity unknown to us) is attempting to do the same. Yet more labs are reproducing other studies, yielding a rich dataset of closely matched reproductions.

1.2 Target paper

Our target paper is Puduppully and Lapata 2021. This paper proposed a neural data-to-text model with a macro-planning stage (determining the high-level organisation of the text-to-be-generated, based on the provided input) followed by a generation stage (where the text is produced). This model is trained and evaluated on both the RotoWire and the MLB datasets (Wiseman et al., 2017; Puduppully et al., 2019). We refer to this model as *Macro*.

The authors carried out an automatic evaluation and two human evaluations. We focus solely on the latter. Experiment 1 asked crowd workers to count supported and contradicting facts in the generated texts (compared to the input data). Experiment 2 asked crowd workers to compare pairs of generated texts in terms of different quality dimensions (discussed in more detail below). In these evaluations, the Macro system was compared to the reference data (referred to as *Gold*), Template-based systems from Wiseman et al. (2017) and Puduppully et al. (2019), ED+CC (again from Wiseman et al.) and Hier (the hierarchical model from Rebuffel et al. 2020, also referred to as RBF-2020 in the original paper). The overall results of these evaluations are highly favourable to the Macro system.

Proceedings of the Third Workshop on Human Evaluation of NLP Systems, pages 75–88, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-088-5_007

1.3 Reproduction target & research question

This paper aims to reproduce Experiment 2 of Puduppully and Lapata (2021). The authors asked crowdworkers to inspect pairs of summaries, and to choose which summary is better in terms of three different quality dimensions (original definitions):

1. **Grammaticality** "Is the summary written in well- formed English?"

2. **Coherence** "Is the summary well structured and well organized and does it have a natural ordering of the facts?"

3. **Conciseness/repetition** "Does the summary avoid unnecessary repetition including whole sentences, facts or phrases?"[1]

The authors used Best-worst scaling (Louviere et al., 2015, BWS) to obtain scores for the three different quality dimensions. In the context of human evaluation of dialogue system output, Santhanam and Shaikh (2019) show that human ratings for coherence and readability are more reliable with magnitude estimation than with BWS. This result was replicated by Braggaar et al. (2022), who obtained similar results. Also in other domains, BWS has been shown to be more reliable than rating scales (e.g. Kiritchenko and Mohammad 2017 for sentiment annotations). In the domain of data-to-text, we are not aware of any studies looking into the reliability of BWS for human evaluations of NLG output. Thus, the main question we aim to answer in this study is: **How reproducible is best-worst scaling for human evaluation of NLG output?**

This question comes with the immediate disclaimer that we are only looking at one implementation of a human evaluation experiment using best-worst scaling, but as noted above: we need to start somewhere. Future studies may alter different parameters of the experiment under consideration, and show if and how these affect the results.

1.4 Contributions

This paper presents a reproduction study answering the research question outlined above. Beyond that, we offer additional analyses of the responses, providing more insight into participant behavior. Finally, we offer reflections on the reproduction process and a proposal for a future study using the data from both reproduction studies targeting experiment 2 of Puduppully and Lapata (2021). Our code and data are available online.[2]

2 Method

Next to the original paper and materials,[3] we also have the support of the original authors. Because multiple labs are all reproducing individual experiments from each paper-to-be-reproduced, we contacted the authors through the coordinator of the ReproHum project, who collated all answers in a shared online document. For the current paper under investigation, this meant that four labs (and the ReproHum coordinator) critically read the paper and asked questions about the methodology. Although this resulted in useful additional documentation, some details about the original study were still missing (as documented below).

Design. We tried to match the original experiment as closely as possible. The original authors used a classical crowdsourcing design, where each ranking decision (indicating which summary is better in terms of a given quality dimension) was distributed as a separate Human Intelligence Task (HIT) on the Mechanical Turk platform. Figure 1 provides an example HIT (without the information letter or informed consent form).

Materials. The original study compared the outputs of four systems with gold-standard summaries generated by humans. For each of the five groups (four systems plus humans), there were 20 summaries. Originally the comparison was made for two separate datasets (MLB and Rotowire), but our reproduction focuses on the outputs for the Rotowire dataset.[6] This means that there are 20 summaries × 10 combinations = 200 items to be

[1] The original authors seem to use the two terms interchangeably in their paper and materials. In the remainder of this paper we use the term *repetition* because *conciseness* is a more general term, typically indicating a preference for brevity while communicating all relevant information.

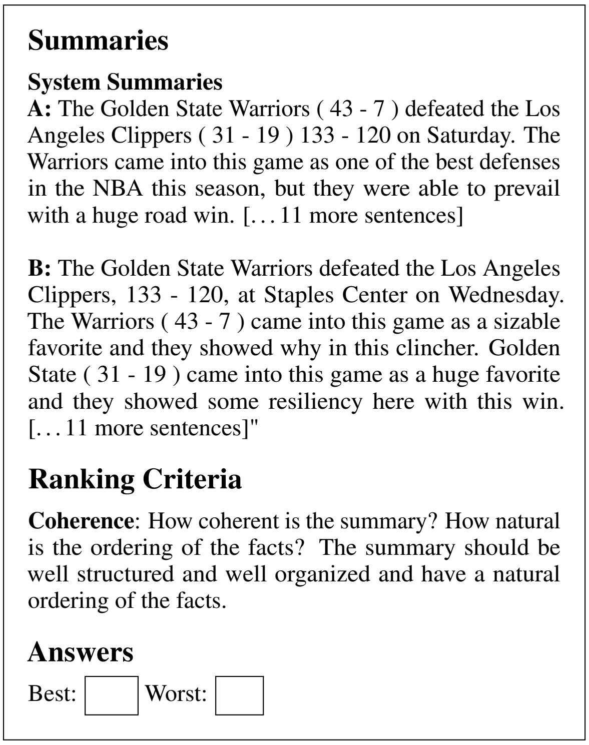
[2] https://github.com/evanmiltenburg/ReproHum-D2T

[3] The lead author of the original paper shared relevant code and data via public GitHub repositories[4,5] and we also obtained the original crowdsourcing templates for use on the Mechanical Turk platform. Details about the evaluation are also provided in the lead author's PhD dissertation (Puduppully, 2022, Appendix B).

[4] https://github.com/ratishsp/data2text-macro-plan-py

[5] https://github.com/ratishsp/data2text-human-evaluation

[6] There was an error in the instructions to prepare the data for the MLB experiment. This error was introduced as the code, data, and instructions were prepared for the ReproHum project and uploaded to GitHub. We do not know what the actual original script looked like. This uncertainty makes the comparison between any replication and the original study unreliable, since we do not know whether the replication corresponds to what was done for the original study. Thus, we leave out the MLB dataset.



Summaries

System Summaries

A: The Golden State Warriors (43 - 7) defeated the Los Angeles Clippers (31 - 19) 133 - 120 on Saturday. The Warriors came into this game as one of the best defenses in the NBA this season, but they were able to prevail with a huge road win. [. . . 11 more sentences]

B: The Golden State Warriors defeated the Los Angeles Clippers, 133 - 120, at Staples Center on Wednesday. The Warriors (43 - 7) came into this game as a sizable favorite and they showed why in this clincher. Golden State (31 - 19) came into this game as a huge favorite and they showed some resiliency here with this win. [. . . 11 more sentences]"

Ranking Criteria

Coherence: How coherent is the summary? How natural is the ordering of the facts? The summary should be well structured and well organized and have a natural ordering of the facts.

Answers

Best: □ Worst: □

Figure 1: Example item showing the ranking task for Coherence. Summaries were manually shortened for presentation in this paper.

rated. With three ratings per item and three quality criteria, there are thus $9 \times 200 = 1800$ ratings to be collected. For these ratings, we use the original HTML interface provided by the authors (see the supplementary materials for the files). This interface contains some (Javascript-based) input validation to ensure that participants can only respond using the characters 'A' or 'B' to indicate their preference.

Participants. Participant location was restricted to English-speaking countries (United States of America, United Kingdom, Canada, Ireland, Australia, or New Zealand). In the general task instructions, participants were told to "attempt HITs if you are a native speaker of English or a near-native speaker who can comfortably comprehend summaries of NBA basketball games written in English." Because of the task's design, for each quality dimension, the participants were able to rate between 1 and 200 items. This also means there is a variable number of unique participants for each quality dimension (see §5.1 for discussion).

Payment. Based on earlier experience with a similar task, we estimate that the average time to complete a HIT would be about 90 seconds. Using a standard wage of $13.59 per hour, this results in a compensation per HIT of $0.34.[7] This is over twice the original amount of $0.15 per HIT, but results from Buhrmester et al. (2011) indicate that compensation level does not seem to influence data quality. A later study from Litman et al. (2015) found similar results for US workers, but noted that greater compensation increased the internal consistency of workers from India. If anything, based on these results we should expect our results to be at least as reliable as the original study.

Procedure. Upon opening the HIT, participants are presented with an information letter and a description of the task. The task description contains the definition of the relevant quality criterion and an example item with an indication of the correct answer. If participants agree to participate, they are asked to provide their informed consent. Having done so, they are presented with two summaries and asked to indicate which summary is the best in terms of the relevant quality criterion. After finishing the HIT, they are optionally asked to indicate whether they are a native speaker of English, and to provide any feedback.

Quality control. Although the original paper made no mention of any quality control measures, these were carried out by the authors. The exact process was not recorded, so our approach is based on the recollection of the authors. This approach was standardised for both concurrent reproductions of the original paper.

For each of the three quality criteria, the HITs were sent out in four batches. The authors used attention checks for two criteria:

[7] Following the ReproHum guidelines, we determined the minimum wage based on Western European standards. We used the maximum of the UK hourly living wage (£10.91 = €12.51)[8] and the standard Dutch minimum wage for a 36-hour workweek (€12.40 per hour).[9] The UK living wage corresponds to $13.59, which is greater than the minimum wages in Canada (CA$16.55 = $12.33), Ireland (€11.30 = $12.24), and more than twice the US minimum wage of $7.55. It is lower than the minimum wages in Australia (AU$ 21.38 = $14.21), New Zealand (NZ$22.70 = $14.17). Thus, the compensation level at least exceeds the median minimum wage for these countries. All wages were taken from government websites. All conversions here are computed using the rates on May 17, 2023.

[8] This amount takes into account the general cost of living in the UK, and exceeds the standard minimum wage. Source: https://www.livingwage.org.uk/what-real-living-wage

[9] The standard differs by sector, depending on the standard amount of hours for one workweek. These hours tend to range between 36 and 40, with fewer hours resulting in a higher wage per hour. Result computed using: https://www.rijksoverheid.nl/onderwerpen/minimumloon/rekenhulp-minimumloon-berekenen

1. *Conciseness* attention check: when considering Gold vs System (excluding template-based systems); if a participant selects a system output with a relatively high amount of repetitions as being more concise than Gold, then it is an exclusion trigger.[10]

2. *Coherence* attention check: when considering Gold vs Template; if a participant selects Template as being more coherent than Gold, then it is an exclusion trigger.

These attention checks were carried out after each batch. No checks were carried out for Grammaticality. The attention checks function as exclusion triggers: failing an attention check means that workers are excluded from working on future batches.[11] Because of the way the crowd sourcing task is set up, not all workers encounter attention checks. So it is possible that low-quality responses remain. Furthermore, following the original authors, we did not publish new HITs to replace the ones that were carried out by workers that were flagged by the exclusion triggers.

Analysis. To determine the inter-rater reliability, we first compute Krippendorff's (2011) alpha for the overall ratings. It is unclear how this was done in the original paper, since there are three different quality dimensions, but only one alpha score was reported. Thus we will report the alpha scores for all three quality dimensions, plus an average of those three values. (Alternatively, one *could* combine all data files for all quality dimensions and compute the overall reliability of participants' preferences, regardless of the relevant quality dimension. However, this misses the point of the alpha score, which is to determine how reliably different constructs can be coded.)

To compare system performance, we use the Best-Worst scaling approach as described in the original paper. For each summary, the output of all systems are compared to each other (for ease of exposition, use of the term 'system' includes Gold responses). This means that each system is compared to four others. For each system, we award a point for every win and we subtract a point for every loss, meaning that for every summary, every system receives a score in the range of [-12,12] (four comparisons per system, times three participants).[12] We use the authors' original scripts to first compute a one-way ANOVA to see if there are any significant differences between the systems, followed by Tukey's HSD to identify which systems differ significantly from each other.

Power analysis. Prior to carrying out our reproduction study, we computed a power analysis to determine the probability to detect a true effect (i.e. finding differences between the systems) if there is one. This turned out to be more difficult than we thought, since the original paper does not report any effect sizes, nor does it report enough information to compute Cohen's d (no standard deviations are reported). Using the available information about the experiment, we estimate that the original experiment had a power of 0.64 to detect a medium-sized effect or greater ($\geq$ 0.3).[13,14] Our study uses the exact same parameters as the original study, and thus has the same power.

3 Results

We first provide some descriptive statistics (§3.1) to contextualise the results, before moving on to the inter-rater reliability (§3.2) and the system comparison (§3.3).

3.1 Descriptives

Table 1 shows the answer frequencies. We find that participants had an overall preference for the first system in the comparisons. Furthermore, despite JavaScript answer validation, some of the respondents provided invalid responses. These are simply

[10] This check was developed by the ReproHum coordinator, to make the original method (relying on human judgments) more reproducible and to keep the process the same between both concurrent reproduction attempts.

[11] We use a *soft block* for this: tagging workers with a custom qualification on Mechanical Turk, and setting a rule that tagged workers cannot take part in our study. This is preferable to a *hard block* (rejecting their work and negatively affecting their performance score) because the rating task is relatively subjective, and a hard block would punish the workers for having the 'wrong' opinion.

[12] Following Orme (2009), the reported scores in the original paper lie between -100 and 100. To obtain scores in this range, we simply carry out a linear transformation of the responses.

[13] We used the pwr library (Champely, 2020) in R (R Core Team, 2023) to run the following command: `pwr.anova.test(k=5,f=.3,sig.level=.05,n=20)` These numbers correspond to the number of different systems (5), desired effect size (0.3 or greater), significance level (0.05), and the number of summaries (20).

[14] One complication in the design of the current study is that it is not straightforward to discuss sample size. There were 206 participants in the original study, but they all provided different numbers of ratings. These were then aggregated to produce the scores for each ⟨system, summary⟩ pair. The reliability of the scores for each system depends on the number of binary judgments per combination of systems. The reliability of the statistical analysis depends on the number of summaries that the systems were evaluated on.

Category	A	B	5	19	Total
Grammaticality	319	277	4	0	600
Repetition	305	287	7	1	600
Coherence	320	277	3	0	600
Total	944	841	14	1	1800

Table 1: Answer frequencies per quality dimension. The answers '5' and '19' are wrongly provided.

skipped in the original best-worst scaling procedure. For other statistics, we do not know how invalid responses were dealt with. We will take up this issue in Section 3.2, when we discuss inter-rater reliability and Krippendorff's Alpha.

Table 2 shows the number of participants in our experiment. Overall, the number of unique respondents (216) is similar to the original experiment (206). We also see that participants carried out HITs for different quality criteria: one participant carried out 67 HITs overall, while the highest number of HITs for any participant on a single quality criterion is 36. We further find that Grammaticality has the lowest number of unique participants, which may be due to the fact that there were no attention checks for this criterion.

Table 3 shows the duration of each HIT. We observe that both mean and median times differ significantly between tasks, but we do not know why.[15] Given the extremely long times taken to complete each HIT, we believe the time to complete each HIT may reflect crowd working strategies of the participants more than they reflect task difficulty.

3.2 Inter-rater reliability

We computed separate Krippendorff's alpha scores for each construct, obtaining a score of α=0.131 for Coherence, α=0.0438 for Grammaticality, and α=0.203 for Repetition. The original authors did not specify (and could not remember) how Krippendorff's alpha was computed, but these were the highest scores after multiple different attempts. We computed Krippendorff's alpha:

1. using a sparse matrix where each row represents a worker and each column represents an item.
2. using a sparse matrix as before, but removing any responses that were not allowed. (For example, one worker responded with '5' while only the values A and B are allowed.) This gave the best result.
3. using a dense matrix where we have three rows representing the first, second, and third response for each item, and each column represents an item. The results from this approach were more or less equivalent to the first approach.

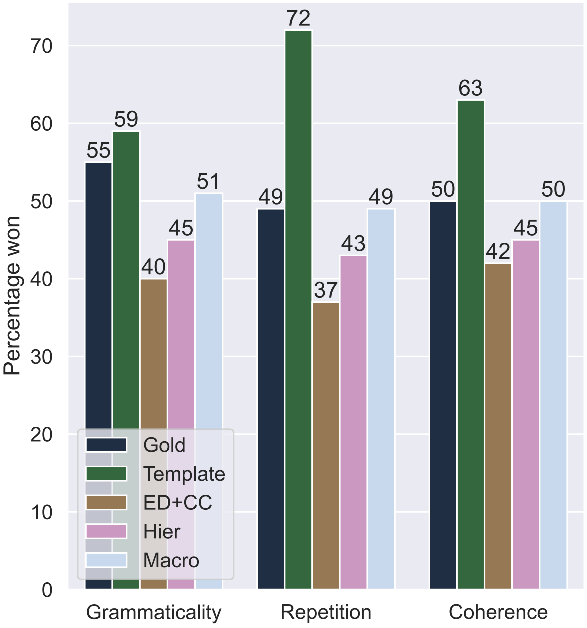


Figure 2: Percentage of system wins across all system comparisons, separated by task. Since we do not have the original data, we cannot compare our results to the original study.

Our results are a far cry from the α=0.47 in the original paper. In Section 4.1 we will further investigate the annotator quality through two different percentage agreement scores.

3.3 System comparison

Figure 2 shows the percentage of system wins across all system comparisons, separated by task. We observe that, using this metric, the template-based approach beats all other systems, including the gold standard summaries. This is surprising, to say the least, since in the original paper the template-based approach is actually the *worst* system across the board.

We now turn to the Best-Worst Scaling (BWS) approach used in the original paper. Given the initial results in Figure 2, it is to be expected that

[15] We initially wanted to run an ANOVA to determine whether there are any significant differences between the three groups. Since this analysis assumes equality of variance, we first ran Levene's test. This test was significant (F(2,1797)=27.91, p<0.05), indicating that this assumption of the ANOVA was not satisifed, so we ran a Kruskal-Wallis test instead of the ANOVA. This test was significant (H(2)=136.26, p<0.05), indicating that the time per HIT differs between groups.

Category	Total	Min	Max	Mean	Stdev	Attention check
Overall	216	1	67	8.33	12.24	Mixed
Coherence	119	1	36	5.04	6.46	Yes
Repetition	135	1	33	4.44	6.37	Yes
Grammaticality	80	1	30	7.50	7.79	No

Table 2: Number of participants in our experiment. Total indicates the total number of unique participants per subset. Min, Max, Mean, and Std refer to the number of HITs per participant. The last column indicates the use of attention checks after each batch of 50 items.

Category	N	Mean	Median	Stdev	Min	Max
Overall	1800	73m49s	49m26s	65m19s	31s	239m56s
Coherence	600	73m26s	46m46s	70m0s	33s	239m56s
Repetition	600	52m36s	30m22s	56m17s	1m3s	234m38s
Grammaticality	600	95m25s	99m9s	61m50s	31s	237m59s

Table 3: Duration of each HIT. Times are cut off at the 4 hour mark, since we indicated that they should be completed within 4 hours. This differs from the 7 hours that were allotted to participants in the original experiment, but we doubt that this would have any effect on the results.

		Grammaticality	Coherence	Repetition
Replication	Gold	9.17	-0.42	-1.67
	Template	17.08	25.42	43.75*
	ED+CC	-19.58	-15.00	-25.83
	Hier	-9.58	-10.42	-14.58
	Macro	2.92	0.42	-1.67
		Grammaticality	Coherence	Repetition
Original	Gold	38.33	46.25*	30.83
	Template	–61.67*	–52.92*	–36.67*
	ED+CC	5.0	–8.33	–4.58
	Hier	13.33	4.58	3.75
	Macro	5.0	10.42	6.67

Table 4: Results using Best-Worst scaling. The asterisk indicates a significant difference between the system and Macro. The *Original* label refers to the original RotoWire results from Puduppully and Lapata (2021).

these results will also be different from the original paper. Table 4 shows that this is indeed the case. Whereas the original paper found multiple systems were significantly different from their system using Macro-planning (indicated by the asterisk), we now only find that the Template-based system is significantly better at avoiding repetitions than the system using Macro-planning. Full details about the statistics are provided in Appendix A.

3.4 Quantifying reproducibility

Now we can ask ourselves: how reproducible are the different measures that we aimed to reproduce? We might paraphrase this question as: how similar are our measures to the original measures of system quality? Given that the result of Best-Worst Scaling is a ranking with relative performance scores, the Spearman correlation is a natural fit.[16] For each of the three quality dimensions, we obtain low (and even negative) correlation values, meaning that our Best-Worst Scaling results do not seem associated with the original scores:

Grammaticality: $\rho = -0.21$
Coherence: $\rho = -0.1$
Repetition: $\rho = -0.05$

See Appendix B for a discussion of the CV* metric to quantify the reproducibility of the current experiment.

4 Additional/alternative analyses

4.1 Annotator quality

Next to Krippendorff's alpha, we can also compute other agreement metrics. For example, we can compute proportions for how often each participant agrees with the majority (i.e., at least 2 out of 3 ratings, for any given item). Table 5 shows the mean agreement for all workers (ranging between 0.72 and 0.74). To compensate for the variation in the number of items that were rated by each participant, we also compute a weighted mean where the agreement scores per participant is weighed

[16] With the caveat that the sample size is very small, leading to a less reliable measure of association.

Category	Mean	Weighted Mean
Coherence	0.72	0.78
Repetition	0.73	0.79
Grammaticality	0.74	0.76

Table 5: Mean agreement and weighted mean agreement of workers with the majority response for each item. The mean is computed based using the scores for all individual workers, even if they only carried out one HIT. The weighted mean multiplies each worker's agreement by the total number of HITs they performed, and divides the sum of all scores by the total number of HITs.

by the number of items rated by that participant. The resulting weighted agreement score is higher (between 0.76 and 0.78).

4.2 Mixed effects analysis

To control for possible random item effects of the individual summaries and to explore the extent to which the order in which the summaries were presented to workers influenced their ratings, we performed an additional generalized linear mixed effects analysis for each of the criteria (Coherence, Grammaticality, Repetition). We used the GLMER function from the *lme4* package in R (version 4.3.1.; R Core Team, 2023; Bates et al., 2015).[17] Since comparisons between Macro and the other systems were the main aim of the original authors, we set Macro as the reference category to which the other systems were compared for all three models. We first constructed a maximal model (Barr et al., 2013) that included a random intercept for *Items* and a random slope for *Order*. We started each criterion analysis by construing a maximal model that included the *System*Order* interaction in the fixed effects structure and a random slope for *Order*. For none of the criteria, the maximal model converged (presumably due to sparsity of the data). After removing the random slope for *Order*, the adjusted models converged. However, Likelihood Ratio Tests that compared the model with *Order* in the fixed effects structure to the random intercept for Summary model showed that adding the order in which the summaries were presented to workers did not improve the models' fit for any of the criteria:

Coherence:	$\chi^2(5) = 8.97$, $p = .110$
Grammaticality:	$\chi^2(5) = 4.87$, $p = .432$
Repetition:[18]	$\chi^2(7) = 6.42$, $p = .491$

In other words: presentation order does not significantly influence the results; there is no evidence for a systematic preference for either the first or the second summary. See Appendix D for further discussion of our mixed effects analysis.

4.3 TrueSkill

Next to best-worst scaling, we also carried out a system comparison using the TrueSkill algorithm (Herbrich et al., 2007).[19] Since the performance of some systems may be very similar and a total ordering would not reflect this, we adopt the practice used in machine translation of presenting a partial ordering into significance clusters established by bootstrap resampling (Sakaguchi et al., 2014). In this case, the TrueSkill algorithm is run 1000 times, producing slightly different rankings each time as pairs of system outputs for comparison are randomly sampled. This way we can determine the range of ranks where each system is placed 95% of the time or more often. Clusters are then formed of systems whose rank ranges overlap.

Figure 3 shows the results. We find that only the Template-based and ED+CC system have non-overlapping confidence intervals, for one criterion, namely *Repetition*. Though robust (because of the bootstrapping procedure), this approach does find fewer differences between the systems than the original approach using an ANOVA and Tukey HSD test.

5 Discussion

5.1 Alternative design

One of the challenges of the design used in the original experiment is that for each quality dimension, the raters individually provided between 1 and 200 ratings. This makes it harder to assess inter-rater reliability, and also means that not all raters were presented with an attention check (providing grounds to exclude raters based on their performance). The design of this study could be improved by using larger sets of items, for example asking each participant to rate 50 items. This would allow us to

[17] We restructured the dataset by items (unique generated summaries) and coded the winning system in the comparison with "1" and the other system with "0", meaning that each HIT was represented by two rows, each focused on one of the two compared systems. We added an extra *Order* column, in which we coded whether the target system was the first (0) or second (1) system in the comparison.

[18] Here, the random slope for Order was included in the bigger model in the comparison.

[19] We used the Python implementation available through PyPI.

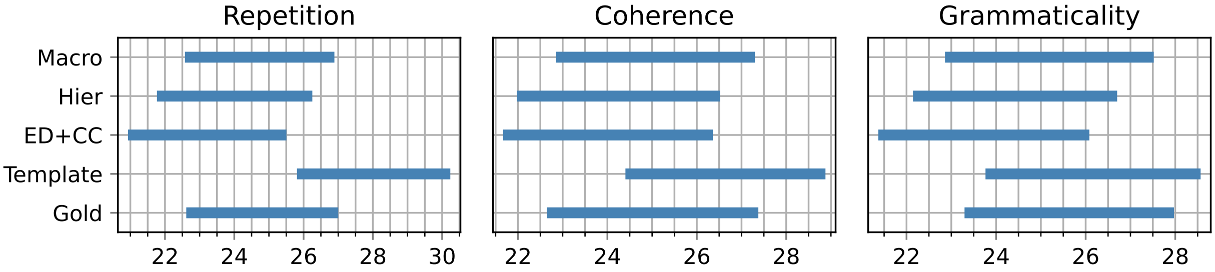


Figure 3: 95%-confidence intervals for the TrueSkill results. When the confidence intervals between two systems do not overlap, we can say that the system outputs are significantly different from each other. This is only the case for the repetition judgments for ed_cc and Template.

validate the performance of each participant, and to assess both inter- and intra-rater reliability.

In the original design, participants rank a pair of summaries but in the end four systems are compared to a gold standard. This is not the only possible implementation of Best-Worst Scaling. For example, Santhanam and Shaikh (2019) asked participants to rank all items at the same time. Presumably the original authors did not do this because the task may have become overwhelming, given the size of the texts. As another option, one could also introduce ties, to indicate that two summaries are roughly of the same quality. Finally, the order of presentation was not randomised in the original study. For each pair of summaries-to-be-assessed, $\langle A, B \rangle$, A was always presented before B.

Alternative design choices may or may not yield more reliable results, but the point is that there is a large parameter space that is ready to be explored. It would be useful for future studies to acknowledge this observation, and to motivate their design choices in more detail. Preregistration may be useful to specify the research methodology early on in the process (van Miltenburg et al., 2021).

5.2 On sample size fidelity

The guidelines for the ReproHum project indicated that we should copy the original set-up as closely as possible, including the number of participants (or in this case: HITs). However, Simonsohn (2015) suggests that the sample size for a replication should be 2.5 times bigger than the sample size estimated for the initial study, to be able to draw reliable conclusions about the reproducibility of the originally observed effects.[20,21] Discussing this idea in full goes beyond the scope of this paper, so for now we simply propose to consider the question: how can we ensure that reproduction studies in NLP provide a reliable estimate of the effects that are demonstrated in the original studies? This question is to some extent complementary to the one posed by Belz (2022): how variable are the human evaluation metrics that are used in NLP/NLG?

5.3 Exceptional circumstances

This reproduction took place in exceptional circumstances, where there were (1) responsive authors (2) who were able to share their original materials, *and* (3) multiple teams of investigators asking critical questions about implementation details for the original study (lowering the chance of overlooking important information, at the expense of time and effort). Thus, our study describes *the best case scenario* for reproduction studies in NLP, which is not representative of reproduction attempts in general. Even in the best scenario, some elements to be reproduced still raise questions. It is now even clearer to us that thorough documentation at publication time is essential, because otherwise many details about the original study may not be recovered.

6 Proposal for follow-up studies

Within the ReproHum project, another lab has simultaneously reproduced the same experiment as

[20] Furthermore, if the difference between systems is truly robust, we should be able to observe the difference through different methods as well. In other words: we might also try to carry out *conceptual* rather than *direct* replications, particularly if the original study is flawed. (See Zwaan et al. 2017; Derksen and Morawski 2022 for a discussion.)

[21] Van Zwet and Goodman (2022) go even further, and argue that the sample size for a replication study should depend on the original p-value. To be able to detect the original effect with high power, one might need a study with a sample size up to sixteen (!) times larger than the original study.

Claim	Reproduced?
Macro is the best system in comparison to the other systems	No
Template is the worst system across the board	No
Multiple systems are significantly different from Macro	No

Table 6: Original claims and their status in our paper.

in this paper. When the data for both experiments are released, this gives us the opportunity to run follow-up studies. Some ideas to consider are: (1) A more in-depth analysis of annotator reliability. (2) A reproduction of the original data analysis using the combined datasets —this at least gets us closer to Simonsohn's proposed sample size for reproduction studies. (3) A simulation study where ratings for the experiment are drawn from a larger pool of ratings and we can determine the amount of variation between different samples. This is similar to the bootstrap resampling strategy we used in the TrueSkill analysis (§4.3), but here we would run the original data analysis multiple times to estimate the range of possible scores for each model using Best-Worst Scaling approach.

7 Conclusion

We carried out a reproduction of Experiment 2 from Puduppully and Lapata (2021), with support from the original authors. We were not able to reproduce the exact results, instead finding opposite trends. For example, the Template-based approach seems to achieve the *best* performance across the board, where it was actually the *worst* performing system in the original paper. (See Table 6 for more.) It is not clear why the results differ from the original study, but we believe that both our study and the original study may be underpowered. Future reproduction studies should probably increase their sample size to make the results more reliable.

Next to the reproduction of the original study, we also provide an extensive selection of descriptive statistics, as well as a set of alternative analyses of the results. With these alternative approaches, we hope to have shown the possibilities and limitations of the experimental design. One key takeaway here is that it is important to have a sufficient amount of ratings per annotator (and ideally the same amount for each annotator). This enables us to dive deeper into the variation within and between ratings from different annotators. Understanding this variation also brings us closer to understanding the replicability of different research results.

Acknowledgments

We would like to thank Craig Thomson for coordinating the ReproHum project and assisting with our study. We also thank Ratish Puduppully for his responses to our inquiries about the original paper. The ReproHum project (and thus our evaluation experiment) is funded by EPSRC grant EP/V05645X/1. Our study was approved by the Research Ethics and Data management Committee (REDC) at the Tilburg School for Humanities and Digital Sciences (TSHD), reference REDC2019.40d.

References

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3).

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Anya Belz. 2022. A Metrological Perspective on Reproducibility in NLP*. *Computational Linguistics*, 48(4):1125–1135.

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.

Anya Belz, Craig Thomson, and Ehud Reiter. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Anouck Braggaar, Frédéric Tomas, Peter Blomsma, Saar Hommes, Nadine Braun, Emiel van Miltenburg, Chris van der Lee, Martijn Goudbeek, and Emiel Krahmer. 2022. A reproduction study of methods for evaluating dialogue system output: Replicating santhanam and shaikh (2019). In *Proceedings of the 15th International Conference on Natural Language*

Generation: Generation Challenges, pages 86–93, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5.

Stephane Champely. 2020. *pwr: Basic Functions for Power Analysis*. R package version 1.3-0.

Maarten Derksen and Jill Morawski. 2022. Kinds of replication: Examining the meanings of "conceptual replication" and "direct replication". *Perspectives on Psychological Science*, 17(5):1490–1505. PMID: 35245130.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. Trueskill(tm): A bayesian skill rating system. In *Advances in Neural Information Processing Systems 20*, pages 569–576. MIT Press.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. Retrieved from https://repository.upenn.edu/asc_papers/43.

Leib Litman, Jonathan Robinson, and Cheskie Rosenzweig. 2015. The relationship between motivation, monetary compensation, and data quality among us- and india-based workers on mechanical turk. *Behavior Research Methods*, 47(2):519–528.

Jordan J Louviere, Terry N Flynn, and A A J Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press, Cambridge.

Bryan Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and hb. Technical report, Sawtooth Software.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with entity modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2023–2035, Florence, Italy. Association for Computational Linguistics.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.

Ratish Surendran Puduppully. 2022. *Data-to-text generation with neural planning*. Ph.D. thesis, The University of Edinburgh.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. A hierarchical model for data-to-text generation. In *Advances in Information Retrieval*, pages 65–80, Cham. Springer International Publishing.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA. Association for Computational Linguistics.

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.

Uri Simonsohn. 2015. Small telescopes: Detectability and the evaluation of replication results. *Psychological science*, 26(5):559–569.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. 2021. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, Online. Association for Computational Linguistics.

Erik W. van Zwet and Steven N. Goodman. 2022. How large should the next study be? predictive power and sample size requirements for replication studies. *Statistics in Medicine*, 41(16):3090–3101.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Rolf A Zwaan, Alexander Etz, Richard E Lucas, and M Brent Donnellan. 2017. Making replication mainstream. *Behavioral and Brain Sciences*, 41:1–61.

A Detailed statistics

We find the following results:

- For Grammaticality, the ANOVA result was significant: F(4,95)=4.027, p=0.005. The Tukey HSD results are provided in Table 7.
- For Coherence, the ANOVA result was significant: F(4,95)=4.313, p=0.003. The Tukey HSD results are provided in Table 8.
- For Repetition, the ANOVA result was significant: F(4,95)=9.802, p<0.001. The Tukey HSD results are provided in Table 9.

Note that, as in the original study, these results were computed before re-scaling the scores to values between -100 and 100.

B Applying CV*

Belz (2022) suggests to use CV* as a general measure of reproducibility, but it is unclear to us whether CV* can or should be applied in this situation. If it can be applied in this case, then we can only compute CV* over two values at a time. For example: comparing the Grammaticality score of a particular system (e.g. Macro) between the original study and our reproduction. With only two data points, the CV* value is probably not very reliable. Having that said, we did run the CV* analysis for completeness' sake.

Because CV* requires all values to be greater than zero, we need to transform the scale from [-100,100] to [0, 200]. For Macro, this results in:
Grammaticality: CV*([102.92, 105])=1.995
Coherence: CV*([100.42, 110.42])=9.457
Repetition: CV*([98.33, 106.67])=8.112
It is not clear how to interpret these values.

C A brief qualitative look at the summaries

The most striking difference we found is the one between template and macro for repetition (as can be seen in Table 4. We take a brief look at two summaries for both macro and template to see if we can detect any patterns that might explain the difference.

The first summary (Figure 4 shows clear repetition in the macro case, but surprisingly only two

System summaries

Template: The San Antonio Spurs (25 - 16) defeated the Portland Trail Blazers (30 - 10) 110 - 96 . LaMarcus Aldridge scored 24 points (9 - 18 FG , 0 - 0 3PT , 6 - 6 FT) to go with 8 rebounds . Damian Lillard scored 23 points (9 - 17 FG , 0 - 5 3PT , 5 - 6 FT) to go with 2 rebounds . Kawhi Leonard scored 20 points (8 - 18 FG , 1 - 4 3PT , 3 - 4 FT) to go with 4 rebounds . Danny Green scored 19 points (6 - 8 FG , 4 - 5 3PT , 3 - 3 FT) to go with 5 rebounds . Patty Mills scored 18 points (7 - 14 FG , 3 - 7 3PT , 1 - 1 FT) to go with 4 rebounds . CJ McCollum scored 17 points (7 - 10 FG , 2 - 2 3PT , 1 - 4 FT) to go with 2 rebounds . The San Antonio Spurs ' next game will be at home against the Dallas Mavericks, while the Portland Trail Blazers will travel to play the Bulls .

Macro: The San Antonio Spurs defeated the Portland Trail Blazers , 110 - 96 , at AT & T Center on Wednesday . The Spurs (25 - 16) came into this game as one of the best defenses in the NBA , but they were able to prevail with a huge road win . The Spurs (30 - 10) have now won five of their last six games , as they continue to slide down the stretch . LaMarcus Aldridge scored a team - high 24 points on 9 - of - 18 shooting to go with eight rebounds , three assists , one steal and one block on 9 - of - 18 shooting . Damian Lillard finished with 23 points on 9 - of - 17 shooting , along with five assists and two rebounds , in 36 minutes . C. J. McCollum chipped in 17 points on 7 - of - 10 shooting , in 20 minutes off the bench . The Trail Blazers shot 50 percent from the field and 32 percent from three - point range . The Spurs shot 45 percent from the field and 38 percent from three - point range . Kawhi Leonard led the way with 20 points , four rebounds , five assists and three steals . Danny Green added 19 points , five rebounds , two assists and two steals , in 34 minutes . Tim Duncan chipped in 11 points , 12 rebounds , six assists , two blocks and one steal , in 30 minutes . Patty Mills chipped in 18 points on 7 - of - 14 shooting , in 22 minutes off the bench . The Spurs shot 45 percent from the field and 38 percent from three - point range , as they shot 45 percent from the field and 38 percent from three - point range . Up next , the Spurs will travel to New Orleans to take on the Pelicans on Friday .

Figure 4: Macro and template example 1.

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	0.95	0.9396	-2.4979	4.3979	False
0	2	-3.45	0.0498	-6.8979	-0.0021	True
0	3	-2.25	0.3713	-5.6979	1.1979	False
0	4	-0.75	0.974	-4.1979	2.6979	False
1	2	-4.4	0.0053	-7.8479	-0.9521	True
1	3	-3.2	0.0821	-6.6479	0.2479	False
1	4	-1.7	0.6475	-5.1479	1.7479	False
2	3	1.2	0.8689	-2.2479	4.6479	False
2	4	2.7	0.1971	-0.7479	6.1479	False
3	4	1.5	0.7457	-1.9479	4.9479	False

Table 7: Grammaticality: Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	3.1	0.1178	-0.4564	6.6564	False
0	2	-1.75	0.6492	-5.3064	1.8064	False
0	3	-1.2	0.8812	-4.7564	2.3564	False
0	4	0.1	1.0	-3.4564	3.6564	False
1	2	-4.85	0.0024	-8.4064	-1.2936	True
1	3	-4.3	0.0096	-7.8564	-0.7436	True
1	4	-3.0	0.1398	-6.5564	0.5564	False
2	3	0.55	0.9928	-3.0064	4.1064	False
2	4	1.85	0.5994	-1.7064	5.4064	False
3	4	1.3	0.8472	-2.2564	4.8564	False

Table 8: Coherence: Multiple Comparison of Means - Tukey HSD, FWER=0.05.

group1	group2	meandiff	p-adj	lower	upper	reject
0	1	5.45	0.0023	1.4621	9.4379	True
0	2	-2.9	0.2635	-6.8879	1.0879	False
0	3	-1.55	0.8159	-5.5379	2.4379	False
0	4	0.0	1.0	-3.9879	3.9879	False
1	2	-8.35	0.0	-12.3379	-4.3621	True
1	3	-7.0	0.0	-10.9879	-3.0121	True
1	4	-5.45	0.0023	-9.4379	-1.4621	True
2	3	1.35	0.88	-2.6379	5.3379	False
2	4	2.9	0.2635	-1.0879	6.8879	False
3	4	1.55	0.8159	-2.4379	5.5379	False

Table 9: Repetition: Multiple Comparison of Means - Tukey HSD, FWER=0.05

System summaries

Template: The Portland Trail Blazers (2 - 2) defeated the Minnesota Timberwolves (2 - 1) 106 - 101 . Damian Lillard scored 34 points (14 - 25 FG , 4 - 9 3PT , 2 - 3 FT) to go with 2 rebounds . Kevin Martin scored 24 points (7 - 12 FG , 2 - 4 3PT , 8 - 11 FT) to go with 2 rebounds . CJ McCollum scored 18 points (7 - 18 FG , 2 - 6 3PT , 2 - 2 FT) to go with 6 rebounds . Al-Farouq Aminu scored 17 points (7 - 12 FG , 2 - 5 3PT , 1 - 2 FT) to go with 9 rebounds . Andrew Wiggins scored 16 points (5 - 17 FG , 0 - 3 3PT , 6 - 7 FT) to go with 6 rebounds . Gorgui Dieng scored 12 points (6 - 9 FG , 0 - 0 3PT , 0 - 1 FT) to go with 5 rebounds . The Portland Trail Blazers ' next game will be at home against the Dallas Mavericks, while the Minnesota Timberwolves will travel to play the Bulls .

Macro: The Portland Trail Blazers (2 - 2) defeated the Minnesota Timberwolves (2 - 1) 106 - 101 on Friday . Damian Lillard had a game - high 34 points on 14 - of - 25 shooting , to go along with seven assists , two rebounds , two steals and one block , in 38 minutes . C. J. McCollum added 18 points , six rebounds , four assists , one steal and one block , in 36 minutes . Al-Farouq Aminu chipped in 17 points , nine rebounds , one assist and one block , in 32 minutes . The Trail Blazers shot 46 percent from the field and 30 percent from three - point range . The Timberwolves , who shot 43 percent from the field and 23 percent from beyond the arc . Kevin Martin led the team in scoring , putting up 24 points on 7 - of - 12 shooting in 35 minutes off the bench . Andrew Wiggins struggled shooting , going 5 - of - 17 from the field and 0 - of - 3 from three - point range . Ricky Rubio added 12 points , nine assists , nine rebounds and three steals in 32 minutes . The Timberwolves shot just 43 percent from the field and 23 percent from three - point range .

Figure 5: Macro and template example 2.

out of three wins are given to template (while template does not show obvious repetitions). In the macro summary there is repetition both between and within sentences ('(...) The Spurs shot 45 percent from the field and 38 percent from three - point range . (...) The Spurs shot 45 percent from the field and 38 percent from three - point range , as they shot 45 percent from the field and 38 percent from three - point range . (...)')

The example in Figure 5 shows no such obvious repetitions. It is clear that macro is quite a bit longer than the summary generated by a template. The template text looks more concise (without fully describing all game statistics, only showing them briefly), focusing more on the key details and briefly describing the next game (which does not happen in macro). In this case template wins three out of three times. Surprisingly not because there are obvious repetitions, but maybe the short text without too many details and only showing the most essential facts is appreciated.

D Further results from the Mixed Effects analysis

We set the probability distribution on binomial with a logit link function and we used parametric bootstrapping over 100 iterations to estimate the confidence intervals and p-values. The complete results can be found in Table 10.

At 95% CI, the results of our mixed effects analyses largely confirm the findings of Section 3.3 in that Macro is significantly different, but worse, for Coherence and Repetition. However, in this analysis, we also find that Macro performs significantly better than Ed+CC for Grammaticality and Repetition.

	System	*B*	*SE* b	99% CI
Coherence	Macro	0.01	0.15	-0.42, 0.39
	Gold	-0.02	0.21	-0.59, 0.53
	Template*	0.53	0.21	0.10, 0.94
	Ed+CC	-0.33	0.19	-0.84, 0.16
	RBF-2020	-0.21	0.22	-0.78, 0.37
	System	*B*	*SE* b	99% CI
Grammat.	Macro	0.03	0.16	-0.36, 0.48
	Gold	0.13	0.21	-0.47, 0.66
	Template	0.31	0.21	-0.28, 0.84
	Ed+CC*	-0.44	0.23	-0.91, -0.009
	RBF-2020	-0.23	0.21	-0.79, 0.29
	System	*B*	*SE* b	99% CI
Repetition	Macro	-0.03	0.17	-0.46, 0.41
	Gold	-0.004	0.25	-0.64, 0.68
	Template**	1.06	0.25	0.46, 1.74
	Ed+CC*	-0.55	0.24	-1.05, -0.08
	RBF-2020	-0.30	0.25	-1.01, 0.33

Table 10: The estimated coefficients and standard errors for the GLMER models that were fitted to workers' ratings of Coherence, Grammaticality, and Repetitio; Macro represents the intercept for all models. Significant at 95% CI = *, at 99% CI = **.