ReproHum: #0033-03: How Reproducible Are Fluency Ratings of Generated Text? A Reproduction of August et al. 2022

Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Martijn Goudbeek, Emiel Krahmer, Chris van der Lee, Steffen Pauws, Frédéric Tomas
Tilburg center for Cognition and Communication, Tilburg University
Warandelaan 2, 5037 AB Tilburg, The Netherlands
✉ C.W.J.vanMiltenburg@tilburguniversity.edu

Abstract

In earlier work, August et al. (2022) evaluated three different Natural Language Generation systems on their ability to generate *fluent*, *relevant*, and *factual* scientific definitions. As part of the ReproHum project (Belz et al., 2023), we carried out a partial reproduction study of their human evaluation procedure, focusing on human fluency ratings. Following the standardised ReproHum procedure, our reproduction study follows the original study as closely as possible, with two raters providing 300 ratings each. In addition to this, we carried out a second study where we collected ratings from eight additional raters and analysed the variability of the ratings. We successfully reproduced the inferential statistics from the original study (i.e. the same hypotheses were supported), albeit with a lower inter-annotator agreement. The remainder of our paper shows significant variation between different raters, raising questions about what it really means to reproduce human evaluation studies.

Keywords: Fluency ratings, Natural Language Generation, Evaluation, Reproduction

1. Introduction

The quality of automatically generated texts is often evaluated using human ratings because they allow us to assess a wide range of different kinds of quality dimensions, ranging from FORM (*grammaticality, fluency*) to CONTENT (*correctness, appropriateness*) and SUITABILITY FOR PARTICULAR PURPOSES (*usability, informativeness*). One major challenge for Natural Language Generation (NLG) research is to properly define and operationalise all the different constructs that one may be interested in (Belz et al., 2020). At the moment, there is a lack of standardisation in the field, leading to terminological confusion. *Fluency* is a major culprit; Howcroft et al. (2020) show how different authors use the same term to refer to *fifteen* different constructs. Clearly the term is open to several different interpretations, which makes it particularly important to clearly define it whenever one wants human evaluators to rate different texts in terms of fluency. Moreover, we may question the reproducibility of any task in which annotators are asked to rate fluency because different interpretations of fluency may lead to different fluency ratings, and thus a less reliable evaluation. Thus we set out to reproduce an earlier study using fluency ratings, and to explore the variability of those ratings.

1.1. Reproduction Target

This paper aims to reproduce the *Fluency* ratings from the human evaluation presented by August et al. (2022). The authors used this evaluation to compare three different systems that produce automatically generated scientific definitions for two different domains: newspapers and journal articles. The study described in the original article did not provide any definition of *Fluency*, but rather relied on examples of fluent and disfluent output. The lack of a definition may lead raters to develop their own idiosyncratic notion of *Fluency*, which may lead to more variation in the ratings.

1.2. ReproHum

This study is part of the broader ReproHum project, where different teams of researchers set out to reproduce several different human evaluation experiments (Belz et al., 2023; Belz and Thomson, 2024). Each study is reproduced at least twice, so another lab is also carrying out a reproduction of the same study as the one reported here. There is no coordination between the two labs, other than the general instructions from the ReproHum coordinator. Following these individual efforts, a meta-analysis will be carried out. This paper explores a technique that may be useful for this meta-analysis, namely equivalence testing (Lakens, 2017).

1.3. Additional Contributions

Next to our reproduction of the fluency evaluation by August et al. (2022), we also collected ratings from eight additional participants. We used these ratings to further study the variability in the behaviour of our raters. So next to the statistics reported in the original article (Krippendorff's alpha for inter-annotator agreement, and independent *t*-tests to compare the different systems) we also present a

Proceedings of the HumEval @LREC-COLING 2024, pages 132–144
Turin, Italy. May 21st, 2024.

mixed-effects model and several descriptive statistics to get a better sense of the factors influencing the ratings provided by our human raters. All of our code and data are available via GitHub.[1,2]

2. Method

Since our study aims to reproduce the original findings from August et al. 2022, we tried to match the original study as closely as possible.

Design The original experiment asked two participants to rate 300 definitions on a four-point scale, ranging from 'not at all fluent' (1) to 'very fluent' (4).

Participants The authors used "two trained annotators" to complete the rating task, one of whom is an author of the original paper. It is not clear what constituted the training or whether the raters were native speakers of English. All we know is that the participants have a background in Natural Language Processing (NLP). The authors do mention that "Neither annotator saw the model generations before evaluation or knew which method had generated each definition."

For our participants, we recruited two PhD candidates from the United Kingdom, working on NLP. Neither participant is a native speaker of English, but they have full professional working proficiency, as is clear from their research. Moreover, both have experience assessing the quality of NLG or Machine Translation output.[3]

Compensation We calculated a fair compensation amount as follows. With 300 items, rating 3 one-sentence definitions for fluency[4] per minute, the task would take about 100 minutes. We rounded this up to two hours to be sure that the time estimate would be feasible. We determined the hourly rate using the standardised ReproHum approach: taking the maximum of the minimum living wage in the UK (£10.90 = €12.62)[5] and the minimum wage in the Netherlands (12.79 euros).[6] Multiplying this by two (hours), we obtain 25.58 euros. Given that we compensate our participants using gift vouchers, which can often only be ordered in multiples of 5 or 10 euros, we rounded this amount up to 30 euros per participant.

Materials The original authors selected three models that performed best overall in their automatic evaluation. These are: DExperts (Liu et al., 2021), GeDi (Krause et al., 2021), and a model proposed by the authors (a fine-tuned BART-model (Lewis et al., 2020) with its definitions reranked by a linear SVM classifier). Using each of these models, the authors automatically generated definitions for 50 terms from the News (which the authors refer to as *low complexity*) and Journal (*high-complexity*) domain. This resulted in $50 \times 2 \times 3 = 300$ definitions for the participants to rate.

Ratings were originally provided through an online interface that is used within the original authors' institution. We used the materials and screenshots available to us to port the experiment to Qualtrics, an online survey platform. We used a Python script to generate the full questionnaire, and provide this script along with instructions on how to set up the experiment on the Qualtrics platform.[7]

We know that the original experiment provided instructions to the participants with some examples of *Fluent* and *Not at all fluent* definitions, but these examples were not available to us. Thus the ReproHum coordinator provided examples so that both reproductions of this study would use the same instructions. These are provided in Appendix A.

The original authors did not specify whether annotators carried out the full task in one sitting or whether it was possible to spread out the work over an extended period of time. We decided to split the task into 10 lists of 30 items, so that our participants could take a break after every list.[8] We used a Python script to determine the (random) distribution of items across lists and the order in which these items were presented. Each item was eventually presented as in Figure 1.

Procedure After receiving a recruitment e-mail (Appendix B), participants could indicate their willingness to participate via e-mail. They were then asked to read the information letter (Appendix C) and consent form (Appendix D), and then received a final briefing (Appendix E) on how to carry out the study before proceeding to the actual study. At the start of the study, the participants first read the study instructions (Appendix A) and proceeded to

[1] Main repository: https://github.com/evanmiltenburg/ReproHum-definition-complexity.

[2] For the Human Evaluation Data Sheet (HEDS), see: https://github.com/nlp-heds/repronlp2024.

[3] Although one of the participants in the original study was an author of the paper, we explicitly opted for non-author participants to not steer the results in any way.

[4] We make the assumption that fluency does not involve judging correctness and other content-related aspects, which might take more time. Participants are just looking at the text at surface level.

[5] Conversion via Oanda.com, 13 October 2023.

[6] Based on a 36 hour work week, via the Dutch government's website: Rijksoverheid.nl.

[7] Some parts of this process cannot be automated (e.g. setting answer requirements and implementing the survey flow). This makes our study harder to reproduce, so readers intending to reproduce our work should precisely follow these steps.

[8] This is also part of our ethical considerations: helping our participants avoid any injuries due to repetitive work.

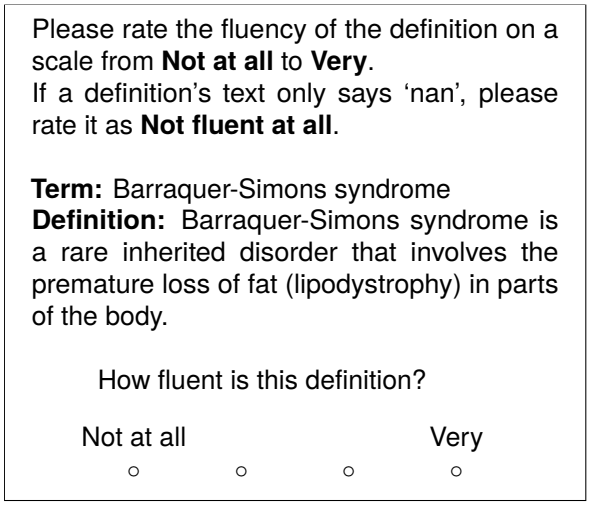

Figure 1: Example item from our study. Note that each item is accompanied by instructions on how to rate it, and the intermediate points are unlabeled.

Model	Original Fluency (SD)	Reproduction Fluency(SD)	CV*
SVM	3.71 (0.59)	3.12 (0.92)	17.225
GeDi	3.20 (1.06)*	2.57 (1.21)*	21.772
DExpert	2.33 (0.85)*	2.28 (1.00)*	2.163
Pearson correlation: 0.95, p=0.208			
Spearman correlation: 1.00, p=0.00			

Table 1: Fluency ratings from the original study and our reproduction. * =Significant compared to SVM ratings using independent *t*-tests corrected for multiple hypothesis testing using the Bonferroni-Holm correction.

rate 10 lists of 30 items. Participants were compensated for their efforts upon completion of the task.

Ethics Approval The format of the original study is very common in NLG evaluation, and does not pose any risk to participants (other than being exposed to flawed definitions of technical terms). The original authors obtained approval from their institution's internal review board to carry out the study. For our reproduction study, we also obtained approval from our local ethics committee[9] before commencing the study.

3. Results

Two participants (with IDs #1 and #2) annotated the data with a Krippendorff's α of 0.52. This is 0.11 lower than the original score of α =0.63. We then ran the same analysis as in the original paper.

[9]The "Research Ethics and Data Management Committee" of the Tilburg School of Humanities and Digital Sciences. Approval code: REDC2019.40e.

3.1. Reproduction Study Results

Table 1 shows the overall results compared to the original study (plus Type I and Type II results, explained below). We find the same pattern as the original paper. The SVM-reranked definitions were rated close to "Very" fluent (3.12 on a 4 point scale), and significantly more fluent compared to GeDi (t_{398} = 5.157, p < 0.001, Cohen's d = 0.516) and DExpert (t_{398} = 8.819, p < 0.001, d = 0.882).

3.2. ReproHum Result Types

The ReproHum project identifies four different kinds of results for a reproduction study. These are:

Type I results The CV* measure (Belz, 2022) is reported to indicate the precision of the evaluation instrument. In other words: the extent to which the measurements vary between different attempts.

Type II results Different correlation measures between the ratings for the different systems. The Spearman correlation shows the extent to which the ordering is the same, while the Pearson correlation shows the extent to which there is a linear relation between the results of the original study and the reproduction.

Type III results Agreement metrics are reported to indicate to what extent annotators in the reproduction study are in agreement with the original annotators.

Type IV results Whether the results of the reproduction study still support the same conclusions as in the original study.

3.3. ReproHum Result Overview

Table 1 shows the CV* values for our reproduction study, as well as the Pearson/Spearman correlations between the original study and the reproduction. These correlations show that, although the means in our study are slightly different, the ranking of the models is the same. However, due to the small sample size (CV* being computed over two scores, and correlations being computed for the scores of three systems) these results should be interpreted with caution.

We are unable to provide Type III results, due to the original ratings being unavailable. Section 5 does provide some more statistics about the inter-annotator agreement.

As for the Type IV results, the original study reported that the SVM-based model has significantly higher scores than both GeDi and DExpert. Our reproduction yields the same conclusion.

	1	2	3	4	5	6	7	8	9	10
1	1	0.65	0.46	0.63	0.61	0.6	0.64	0.49	0.64	0.7
2	0.65	1	0.45	0.38	0.47	0.6	0.71	0.39	0.47	0.52
3	0.46	0.45	1	0.55	0.67	0.68	0.64	0.7	0.74	0.59
4	0.63	0.38	0.55	1	0.65	0.57	0.57	0.63	0.68	0.63
5	0.61	0.47	0.67	0.65	1	0.79	0.71	0.77	0.79	0.76
6	0.6	0.6	0.68	0.57	0.79	1	0.78	0.77	0.84	0.76
7	0.64	0.71	0.64	0.57	0.71	0.78	1	0.72	0.77	0.74
8	0.49	0.39	0.7	0.63	0.77	0.77	0.72	1	0.76	0.73
9	0.64	0.47	0.74	0.68	0.79	0.84	0.77	0.76	1	0.79
10	0.7	0.52	0.59	0.63	0.76	0.76	0.74	0.73	0.79	1

Figure 2: Spearman correlations between all participants based on their available ratings. Participants 1 and 2 (in the 'official' reproduction) and 9 and 10 (in our internal reproduction) provided 300 ratings, while the others rated 120 items.

4. Additional Study

We set out to further explore the variability in fluency ratings. This required us to collect additional ratings so we could compare different raters with each other and establish the range of possible (dis)agreement between them.

Participants All eight authors of this paper provided additional ratings through the same interface. Although all participants are fluent in English and are familiar with the field of Natural Language Generation, none of the participants are native speakers of English. Because we rated the items ourselves, no further compensation was necessary.

Procedure The participants were asked via email to complete four out of ten lists of 30 items, for a total of 120 items per participant. All participants were assigned a numerical identifier (ranging from 003 to 010) so that they could provide their responses anonymously. They then followed the same procedure as in the participants in the base experiment. Two participants volunteered to complete all ten lists of 30 items, for a total of 300 items per participant. As we will see later, this enables us to reproduce our reproduction study.

5. Additional Results

5.1. Variation between Different Raters

For this study, we set out to explore the range of variation between all ten participants (two independent raters, plus eight authors). Figure 2 shows the Spearman correlation between all of our participants. These values range between 0.38 (a low correlation) and 0.84 (a high correlation). For each rater we also computed the average correlation with the other raters. These values range between 0.47 (low) to 0.65 (moderate). We also computed Krippendorff's α over all raters, which resulted in a score of 0.55. This score does not exceed the threshold value of 0.67 that is commonly deemed good enough to draw tentative conclusions (Artstein and Poesio, 2008).

At first glance, it seems unfortunate that the rater with the poorest average correlation score (rater 2) was part of our 'official' reproduction study.[10] Still we managed to reproduce results from the original study, suggesting that the difference between the systems was fairly stark to appear despite the noisy ratings. This clear difference is also reflected in the original effect sizes of 0.6 (medium) for the comparison of SVM-RERANK with GEDI, and 1.88 (very large) for the comparison with DEXPERT.

5.2. Score Distribution

Figure 3 shows the distribution of the scores we obtained in our study. We observe that there is a clear gap between the DEXPERT model and the other two approaches, which both perform much better. The SVM-RERANK model also outperforms GEDI, albeit by a smaller margin. These results mirror the ones from our reproduction in the previous section. For future studies in this area, one might wonder whether a four-point scale is distinctive enough, given that over 80% of the scores for the state-of-the-art system (SVM-RERANK-*) have a score of either 3 (over 25%) or 4 (over 50%). Direct Assessment (Graham et al., 2017) may be preferable to tease newer systems apart.

5.3. Duration

Table 2 shows the time each participant spent on a single list of 30 items, rounded to the nearest minute. The median[11] time for one list is about seven minutes, which means that they spent about fourteen seconds on each item. When we extrapolate this to all 300 items, a typical participant would spend about an hour and ten minutes on the full task. This is half an hour faster than our original estimate, and fifty minutes faster than the two hours that we used to determine a fair compensation for this task. (Of course this ignores any overhead costs, such as communication with the study coordinator, startup time, and so on.)

[10] The poor correlation with other raters may not be due to a poor performance. This rater accidentally rated three lists twice, enabling us to measure their consistency between different attempts. This yielded a score of 0.85, meaning that their scoring behavior at least seems internally consistent, and not random.

[11] The median is used because it is less sensitive to outliers (that is: unusually high values), which usually are the result of leaving the form open in the background and completing it later.

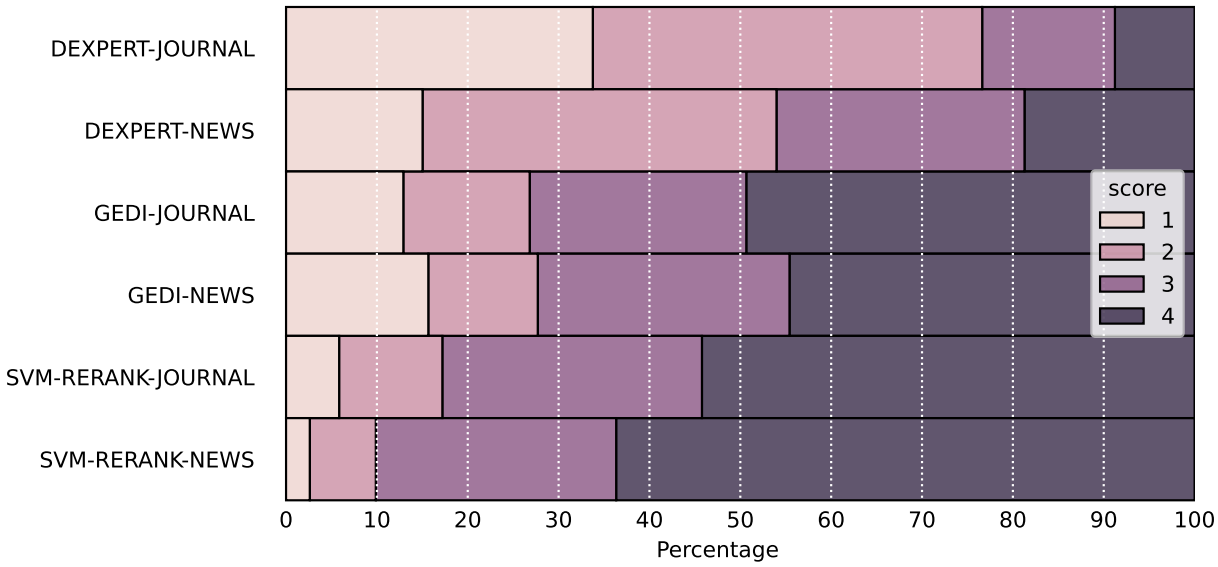

Figure 3: Score distribution for the different models, split by definition complexity (using all scores that we have collected, from all 10 raters). Wider bars indicate a greater proportion. A score of 1 means that a definition is *Not at all fluent* while 4 means *Very fluent*. Thus we find that the DExpert model scores lower (i.e. has fewer scores of 3 and 4) than GeDi, which in turn scores lower than SVM-Rerank.

ppt	min	max	mean	med	std	total
1	21	1284	195	48	388	1946
2	3	42	11	9	11	113
9	4	8	6	6	1	55
10	3	32	8	5	9	76
3	3	10	5	3	4	19
4	4	1421	359	5	708	1437
5	7	11	9	10	2	37
6	6	17	10	9	5	40
7	4	35	16	11	14	62
8	4	12	7	5	4	26
Overall	3	1421	60	7	237	3812

Table 2: Time (in minutes) spent per list of 30 items, by each participant and overall. Abbreviations: ppt=participant, min=minimum, max=maximum, med=median, std=standard deviation. Only participants 1, 2, 9, and 10 (top part of the table) carried out the full task. The other raters (bottom) only scored four lists of items.

5.4. Reproducing our Reproduction

Because two of our participants rated all items, we can also reproduce our reproduction. Our goal here is twofold: first we wish to see whether we obtain the same significant differences between SVM-rerank and GeDi/DExpert. Second, if we find a similar result to our reproduction, then we wish to test the hypothesis that there is no significant difference between the mean system ratings for participants 1&2 versus participants 9&10.

Table 3 and 4 show our results. As with our first reproduction, the SVM-reranked definitions were rated close to "Very" fluent (3.62 on a 4 point scale), and significantly more fluent compared to GeDi (t_{398} = 4.903, p < 0.001, Cohen's d = 0.490) and DExpert (t_{398} = 17.155, p < 0.001, d = 1.716). We did find that our second reproduction achieves mean scores that are much closer to the original study. The effect size for the difference between SVM and DExpert is also much closer between the original study and our second reproduction.

Since we find similar significant results, we can test whether both our reproductions yield scores that are not significantly different from each other. For this, we first used an equivalence test (Lakens, 2017) with the null hypothesis that the effect size of the difference between the two sets of scores is larger than our smallest effect size of interest (sesoi), which we set to 0.2 (the smallest detectable effect, with the bounds set as $\Delta_{Low} = -0.185, \Delta_{Upp} = 0.185$).[12,13] We failed to reject this hypothesis (p=1.00), meaning that we cannot reject the null hypothesis that there is a true effect that is

[12] Lakens (2017) note that one objective way to determine the sesoi is to find "the smallest observed effect size that could have been statistically significant in a previous study." For this we can compute the critical t-value in R: `ct = qt(p=.05/2, df=398, lower.tail=FALSE)`. We can then determine the smallest significant effect: `d = ct * sqrt((1/200) + (1/200))` (where 200 is the sample size for each group –100 judgments per model per rater, for 2 raters). This yields an effect size of 0.2.

[13] These bounds correspond to the maximum difference between the means (Δ). Two one-sided tests are carried out to determine if $\Delta \leq \Delta_{low}$ or $\Delta \geq \Delta_{upp}$.

Model	Original Fluency (SD)	Reproduction 1 Fluency(SD)	Reproduction 2 Fluency(SD)	△OR1	△OR2	△R1R2
SVM	3.71 (0.59)	3.12 (0.92)	3.62 (0.64)	0.59	0.09	0.50
GeDi	3.20 (1.06)*	2.57 (1.21)*	3.23* (0.94)	0.23	0.03	0.34
DExpert	2.33 (0.85)*	2.28 (1.00)*	2.27* (0.92)	0.05	0.06	0.01

Table 3: Fluency ratings from the original study and both our first and second reproductions. * =Significant compared to SVM ratings using independent *t*-tests corrected for multiple hypothesis testing using the Bonferroni-Holm correction. Delta indicates the absolute difference between the Original result (O) and the first reproduction (R1), the original result and the second reproduction (R2), and both reproductions.

	Original	R1	R2
SVM versus GeDi	0.60	0.52	0.49
SVM versus DExpert	1.88	0.88	1.72

Table 4: Effect sizes (Cohen's D) from the original study, the first reproduction (R1) and the second reproduction (R2).

	Est.	SE	*t*-value	95% CI
(Intercept)	2.27	0.11	21.58	[2.06, 2.49]
GeDi	0.80	0.05	15.67	[0.70, 0.90]
SVM-Rerank	1.18	0.05	23.23	[1.08, 1.27]
Category: Wiki	-0.25	0.04	-5.90	[-0.33, -0.17]
Domain: News	0.20	0.04	4.93	[0.12, 0.29]

Table 5: Estimates (Est.), standard error (SE), *t*-values, and 95% confidence interval (95% CI) for the fixed effects.

at least as big as the SESOI. A follow-up analysis revealed a significant difference between our two reproductions (t_{398} = -6.299, p < 0.001, d = -0.63; a medium-sized effect).

Our results show that while both reproductions show the same patterns, and thus support the original claims about the relative performance of the different systems, we cannot reproduce the absolute ratings; different participants use the fluency rating scale differently (but consistently so).

5.5. Mixed-effects Analysis

We also carried out a mixed-effects analysis of the data, incorporating different factors that might influence the ratings. We used the `lme4` library in R (Bates et al., 2015; R Core Team, 2023) to fit a linear mixed effect model with model type (DExperts/GeDi/SVM-Rerank) and domain (news/journalism) as fixed effects. Participant was added as a random effect. Variance at the participant level was 0.09 (*SD* = 0.30).

The results of the fixed effects can be found in Table 5. The 95% confidence intervals show that all of these variables explain to some extent the ratings that were given. More specifically, these results show that both GeDi and SVM outperform DExpert; the models generally perform worse for terms and definitions collected from Wikipedia science glossaries (as compared to MedQuAD); and that the models that were trained using scientific news articles generally perform better than the ones trained using scientific abstracts.[14]

6. Omissions and their Consequences

We successfully reproduced the fluency evaluation from August et al. (2022). With the original paper and some additional information from the authors, it was possible to reproduce the original study, but there were still some omissions, listed below.

Annotators Demographic information about the annotators was incomplete. It is unclear how the annotators were trained. Future authors may wish to use guidelines established by, *inter alia*, Bender and Friedman (2018) or Shimorina and Belz (2022).

Data The definitions used for the experiment were not in the repository associated with the paper, but they were shared by the authors upon request. The raw data for the human evaluation are not available, so we cannot actually see the scores provided by the annotators. This makes it harder to compare our results to the ones in the original paper, and it prevents us from checking for any errors in the statistical analysis. We urge readers to share as much data about their experiments as possible, given the low reliability of data sharing 'upon request' (Krawczyk and Reuben, 2012; Tedersoo et al., 2021; Hussey, 2023).

Procedure The paper does not specify whether the annotation task could be carried out in batches, or whether all 300 items had to be labeled in one single session. For the fluency evaluation, the authors provided the original question, but not the examples that were used to illustrate fluent and

[14] A post-hoc analysis reveals that all models are significantly different from each other, at p <0.0001. (Multiple Comparisons of Means: Tukey Contrasts, with p-values adjusted through the Holm-Bonferroni method. See our GitHub for implementation details.)

non-fluent responses. We also do not know in what order the items were presented to the annotators or whether there was any randomisation involved.

We were happy to see that we managed to reproduce the original results, but what if we had not been able to do so? If it is unclear what the original authors did exactly, it is impossible to pinpoint what deviations from the original procedure could have influenced the results.

Code Although the code for the models is available, there is no code to sample the outputs from the test set and prepare the experiment. The code for the statistical analyses of the human evaluation was also not provided.

Researchers are not infallible. Analytical mistakes are one of the most common sources of error in the retracted scientific literature (Casadevall et al. 2014; also see the Statistics category on the Retraction Watch website). Although there are automatic tools to flag statistical reporting errors (e.g., Nuijten et al., 2016; Brown and Heathers, 2017), having the data and the code used for any statistical analysis is essential to be able to check whether a reported analysis is actually correct.

7. Discussion

7.1. Interpreting reproduction studies

Now that we have reproduced the original study by August et al. (2022), what do our results *mean*? There seem to be at least three different interpretations of the purpose of a reproduction study:

1. In terms of the *hypotheses*: do we find (a lack of) support for the same hypotheses as in the original study?

2. In terms of the *mean*: to what extent do our results differ from the originally reported means? What would the True Means look like?

3. In terms of the *effect size*: regardless of the mean, to what extent does the relative difference between the means differ from the effect sizes reported in the original study? What would the True Effect Size look like?

Whether we have really succeeded in our reproduction depends on which of these interpretations you choose. We have definitely met the first condition: our results provide support for the hypothesis that the SVM-based model has significantly higher scores than both GEDI and DEXPERT. With regard to the second interpretation, we did not successfully reproduce the original study: although the ordering of the system scores is the same, the absolute values we obtained differ quite a bit from the original study.[15] Finally, we also failed to reproduce the original study in terms of the effect size: the original effect size for the comparison between SVM and DEXPERT is twice as large as the one we found in our reproduction.[16]

These questions echo an earlier discussion by Zwaan et al. (2018, particularly §5.6). Our current stance is that the first interpretation of reproducibility is most meaningful in the context of the ReproHum project. If we reproduce an earlier evaluation study, we are mostly interested to see which system performs better. As long as the ordering of the systems is the same, we are happy because we know which NLG techniques tend to work better than others.[17,18]

7.2. Reflections on Fluency

Different raters provided some observations that guided their rating behavior.

7.2.1. Some examples

One rater identified three related but different cases that they treated differently in their ratings.

Case 1: Fluent but uninformative
Term: Heart Valve Diseases
Definition: Your heart is the largest organ inside your body.

Case 2: Fluent but wrong
Term: Salivary Gland Disorders
Definition: Your salivary glands are two small glands in your mouth, each about the size of a fist.

Case 3: Fluent but unhelpful
Term: etchplain
Definition: See etchplain.

This rater argued that the third case is just cheating the system, and marked the system down for it,

[15] One might conclusively (dis)prove this kind of reproducibility through an equivalence test. Even though we do not have the data from the original study, we do have the mean, standard deviation, and sample size. This is enough to run the TOST-procedure.

[16] For less obvious differences, one might compute confidence intervals (CIs) to compare the differences between two effect sizes (Kirby and Gerlanc, 2013; Goulet-Pelletier and Cousineau, 2018; Ben-Shachar et al., 2020). If the CIs overlap, the effect sizes are consistent with each other.

[17] Or when we group systems in different equivalence classes and the ordering of those classes is the same.

[18] Of course, the experimental design should also be controlled enough to be able to learn something meaningful about the performance of NLG systems.

while other raters stuck to a more strict definition of Fluency where the third case was not penalised. This highlights the importance of clear task definitions and clear instructions for raters (as is also recommended by van der Lee et al. (2021)).

7.2.2. A Taxonomy of Errors

Another rater provided a taxonomy of different kinds of issues with the outputs:

- Typos or spelling mistakes e.g., changing the names of medical term in definition, incorrect abbreviation, jumbled two or more words with no meaning.
- Incomplete sentence
- Repetition of specific word
- Minor grammatical errors affecting the naturalness e.g., “electrical” is the right word instead of “electric”, “into” is the right word instead of “to”.
- Sentence structure: having a heading at the beginning of a definition that was not needed e.g., “Summary:”, “Espanol:”.
- Content problems: the given definition did not specifically mention about the disorder/syndrome, or the specific type stated in the term. It only described the location of that gland or heart valves and their generic purpose.
- Relevance: in some cases, it was evident that some definitions had accuracy issues, for example: ‘47,XYY syndrome is a chromosomal condition that affects females. This condition affects “males” but not “females”.’

The ratings that people provide may depend on the perceived severity of these different kinds of errors. Raters may or may not share the same sense of severity for these error categories. (Also see van Miltenburg et al. 2020 for discussion.) One solution to this problem might be to carry out an error analysis rather than rating each output (van Miltenburg et al., 2021). We may also take inspiration from the Multidimensional Quality Metrics (MQM) framework that is used in Machine Translation (Lommel et al., 2013, 2014; Freitag et al., 2021).

7.2.3. Background Knowledge

The same rater observed that (a lack of) background knowledge was an issue for this task, as it is difficult for people without a medical background to understand the fluency of medical terms. For example:

> “Paget disease of bone is a bone disease characterized by abnormal osteoclasts that are large, multinucleated, and overactive and that contain paramyxovirus-like nuclear inclusions.”

The rater indicated that they “do not understand these terminologies but marked this as *very fluent* because it defined the disease and their specific characteristics. Geographic and basic science related terms were comparatively easier.”

Of course, there may also be individual differences in terms of background knowledge, making medical definitions easier to read for some raters than for others. The effort required to read these kinds of texts may also influence rating behavior.

7.2.4. Understanding Variation in Scores

Due to time constraints we were not able to further analyse the results. Still we would like to highlight another way to analyse the data: ranking all items by the extent to which annotators disagree about the score. Metrics to do this include (i) the largest difference between annotators and (ii) the mean squared error of the different scores at the item level. After ranking the items, one could qualitatively analyse the items with the greatest diversity in scores, to identify patterns in the data and develop explanations for variation in annotator behavior.

8. Limitations

Sample size The ReproHum project uses sample size as a control variable, meaning that some reproduction studies (including this one) are required to have the exact same sample size as the original studies that they aim to reproduce. As has been discussed in earlier studies (e.g., van Miltenburg et al. 2023), this limits the power of our reproduction. If we want to know whether a particular instrument (e.g., a rating task) is reliable, we should test it with a larger sample than the original study. We have addressed this issue to some extent, by collecting ratings from eight additional participants and studying the variation in their ratings. However, in terms of participants this is still a small sample size. (It is unclear what would be a good sample size for the outputs that participants are asked to rate.)

Variation due to selected outputs We might also wonder to what extent the assessment of the quality of the systems from the original paper depends on the exact outputs that were selected for the rating task. What would the performance of the systems look like with a different sample of outputs? This is a question that we cannot study, due

to the original outputs being unavailable.[19]

9. Conclusion

We set out to reproduce the study of August et al. (2022) and to explore different factors influencing the variability in Fluency ratings. We followed the original study as closely as possible, with minor inevitable deviations due to some missing information. The results of this reproduction show similar patterns as in the original study, showing significant differences in fluency ratings between the SVM-model and GeDi and between SVM and DExpert. In terms of inter-annotator agreement we found a lower Krippendorff's alpha (0.11 lower) than in the original study. Whether our reproduction is successful depends on your measure of success. Either way, we hope that our statistical *deep dive* into our own reproduction attempt is useful to others wanting to compare the results of different sets of annotators.

10. Acknowledgments

Anouck Braggaar is supported by the Dutch Research Council (NWO) through the *Smooth Operators* project (KIVI.2019.009). We thank Craig Thomson for coordinating the reproduction efforts, our annotators at the University of Aberdeen and Dublin City University, and two anomymous reviewers.

11. Bibliographical References

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

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.

[19]Although it might be possible to re-generate the outputs, there is no guarantee that these will be the same as in the original study, and this would take much more effort than if the original outputs were just directly available.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2024. The 2024 repronlp shared task on reproducibility of evaluations in nlp: Overview and results. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 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.

Mattan S. Ben-Shachar, Daniel Lüdecke, and Dominique Makowski. 2020. effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, 5(56):2815.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Nicholas J. L. Brown and James A. J. Heathers. 2017. The grim test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, 8(4):363–369.

Arturo Casadevall, R Grant Steen, and Ferric C Fang. 2014. Sources of error in the retracted scientific literature. *FASEB J*, 28(9):3847–3855.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang

Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Jean-Christophe Goulet-Pelletier and Denis Cousineau. 2018. A review of effect sizes and their confidence intervals, part i: The cohen's d family. *The Quantitative Methods for Psychology*, 14(4):242–265.

Yvette Graham, Qingsong Ma, Timothy Baldwin, Qun Liu, Carla Parra, and Carolina Scarton. 2017. Improving evaluation of document-level machine translation quality estimation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 356–361, Valencia, Spain. Association for Computational Linguistics.

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.

Ian Hussey. 2023. Data is not available upon request.

Kris N. Kirby and Daniel Gerlanc. 2013. Bootes: An r package for bootstrap confidence intervals on effect sizes. *Behavior Research Methods*, 45(4):905–927.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michal Krawczyk and Ernesto Reuben. 2012. (un)available upon request: Field experiment on researchers' willingness to share supplementary materials. *Accountability in Research*, 19(3):175–186. PMID: 22686633.

Daniël Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Soc Psychol Personal Sci*, 8(4):355–362.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica tecnologies de la traducció*, 12:455–463.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Michèle B. Nuijten, Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, and Jelte M. Wicherts. 2016. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4):1205–1226.

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

Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

Leho Tedersoo, Rainer Küngas, Ester Oras, Kajar Köster, Helen Eenmaa, Äli Leijen, Margus Pedaste, Marju Raju, Anastasiya Astapova, Heli Lukner, Karin Kogermann, and Tuul Sepp. 2021. Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, 8(1):192.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current

trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Debby Damen, Martijn Goudbeek, Chris van der Lee, Frédéric Tomas, and Emiel Krahmer. 2023. How reproducible is best-worst scaling for human evaluation? a reproduction of 'data-to-text generation with macro planning'. In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 75–88, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. Underreporting of errors in NLG output, and what to do about it. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Emiel van Miltenburg, Wei-Ting Lu, Emiel Krahmer, Albert Gatt, Guanyi Chen, Lin Li, and Kees van Deemter. 2020. Gradations of error severity in automatic image descriptions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 398–411, Dublin, Ireland. Association for Computational Linguistics.

Rolf A. Zwaan, Alexander Etz, Richard E. Lucas, and M. Brent Donnellan. 2018. Making replication mainstream. *Behavioral and Brain Sciences*, 41:e120.

A. Instructions for the experiment

You will be given 30 terms with their definitions and asked to rate how fluent the definitions are. You will be asked to rate how fluent the definition is on a scale from **Not at all** to **Very**.

Examples of very fluent definitions:
Term: Acanthoma
Definition: An acanthoma is a skin neoplasm composed of squamous or epidermal cells. It is located in the prickle cell layer.

Term: Transformer
Definition: The Transformer is a deep learning model architecture relying entirely on an attention mechanism to draw global dependencies between input and output.

Examples of not at all fluent definitions:
Term: Acanthoma
Definition: Broad Line Region.

Term: Transformer
Definition: Transformer attention rely.

B. Recruitment e-mail

Dear all,

As part of the ReproHum project at the University of Aberdeen (PI: Prof. Anya Belz, Co-I: Prof. Ehud Reiter), researchers at Tilburg University are looking for two participants to take part in an evaluation of Natural Language Processing (NLP) system outputs.

Participants should be non-student researchers and/or PhD students with some experience in NLP. They should be proficient in English but do not have to be native speakers.

The task is to read 300 definitions that have been produced by different automatic systems and to judge the fluency of those texts. The texts are split up into smaller batches. Since the definitions are short, and fluency is a relatively superficial property (no need to check for factuality), we expect this to take about 2 hours in total. This makes it possible to rate the definitions in between jobs (e.g. while your code is compiling). You will be compensated for your efforts through a €30 gift card.

If you are interested in taking part in this study, please contact Prof. Emiel van Miltenburg by email: C.W.J.vanMiltenburg@tilburguniversity.edu

Thank you,
Craig

C. Information letter

Evaluating the fluency of automatically generated definitions

We invite you to take part in a study on automatic definition generation, carried out by researchers from Tilburg University. Your task is to read 300 definitions which have been produced by different automatic systems, and to judge the fluency of those texts. This enables us to understand which system is best.

Expected duration: there are 300 definitions, split up into 10 batches of 30 definitions. Since the definitions are short, and fluency is a relatively superficial property (no need to check for factuality), we expect this to take about 2 hours in total

(or about 12 minutes per batch). This makes it possible to rate the definitions in between jobs (e.g. while your code is compiling).

We are not aware of any negative consequences to your participation in this task, but please be aware that there may be occasional errors in the generated texts. You will be compensated for your efforts through a €30 gift card.

We remind you that participation is voluntary. You have the right to decline to participate and withdraw from the research once participation has begun, without any negative consequences, and without providing any explanation.

We will not collect any personal data, beyond your general qualification to participate ("a PhD candidate at X university"). We aim to publish the data and results of this study, making your responses publicly available for future research for an indefinite period of time. However, we will ensure that any potentially identifying information (including your IP address, platform ID) will be removed from the data before it is published. Thus, everything will be fully anonymous.

If you have any questions about this study, feel free to contact Emiel Van Miltenburg (C.W.J.vanMiltenburg@tilburguniversity.edu).

This study was approved by the Research Ethics and Data Management Committee (REDC) at Tilburg University (reference: REDC2019.40e). If you have any remarks or complaints regarding this research, you may also contact the "Research Ethics and Data Management Committee" of Tilburg School of Humanities and Digital Sciences via tshd.redc@tilburguniversity.edu

D. Informed consent

Evaluating the fluency of automatically generated definitions

If you would like to continue with this study, please confirm that you have read the information letter and agree with the following terms:

- I have read the information letter.
- I confirm that there was room to ask questions (via email).
- I understand that participation is voluntary.
- I understand that I have the right to decline to participate and withdraw from the research once participation has begun, without any negative consequences, and without providing any explanation.
- I understand and agree that the (anonymised) results from this study will be made publicly available, for an indefinite period of time.
- I agree to participate in this study.

E. Instructions via email

Dear NAME,

Thank you for agreeing to participate in our evaluation study. We will now proceed to the actual task.

Design
As you know, the goal of this task is to rate 300 items. The entire study has been implemented as a survey in Qualtrics, with 10 lists of 30 items. The idea is that you fill in the survey 10 times, one time for each list of items. (With the opportunity to take breaks in between.)

Procedure

1. You can start the task by clicking on the link to the study, at the bottom of this message.
2. A screen with two questions will appear:
   (a) You will be asked for a participant ID. Please fill in your ID: IDENTIFIER.
   (b) You will be asked what set of items you would like to work on. Please complete the task in order. That is: starting with list number 1, and then moving on to list number 2, and so on.
3. The next page provides the full instructions for the task. Please read them carefully.
4. Proceed to rate the 30 items on the list that you have selected.
5. If you are done with the current list of items, you may continue with the next list. This does require you to visit the link to the study again, and to fill in the participant ID again.
6. If you are done with the full task, please send me a message and I will order the gift card based on your instructions. (I.e. where to buy it and where to send it.)

Link to the study: URL.

Final note
I am not sure if Qualtrics allows you to carry out

the same study twice. If not, you can use a private browser window. I have set up the study such that no IP address or any other personal information will be collected.

Thanks again for your participation! Please let me know if you have any further questions.