# The 2022 ReproGen Shared Task on Reproducibility of Evaluations in NLG: Overview and Results

**Anya Belz**
ADAPT/DCU, Ireland and Univ. of Aberdeen
anya.belz@adaptcentre.ie

**Anastasia Shimorina**
Orange, Lannion, France
anastasia.shimorina@orange.com

**Maja Popović**
ADAPT/DCU, Ireland
maja.popovic@adaptcentre.ie

**Ehud Reiter**
University of Aberdeen, UK
e.reiter@abdn.ac.uk

## Abstract

Against a background of growing interest in reproducibility in NLP and ML, and as part of an ongoing research programme designed to develop theory and practice of reproducibility assessment in NLP, we organised the second shared task on reproducibility of evaluations in NLG, ReproGen 2022. This paper describes the shared task, summarises results from the reproduction studies submitted, and provides further comparative analysis of the results. Out of six initial team registrations, we received submissions from five teams. Meta-analysis of the five reproduction studies revealed varying degrees of reproducibility, and allowed further tentative conclusions about what types of evaluation tend to have better reproducibility.

## 1 Introduction

Interest in reproducibility continues to grow across Natural Language Processing (NLP).[1] However, we still do not understand well enough what makes evaluations easier or harder to reproduce, and reproduction studies often reveal alarmingly low degrees of reproducibility not only for human evaluations but also for automatically computed metrics.

With the ReproGen shared task on Reproducibility of Evaluations in NLG, our aim is to add to the body of reproduction studies in order to increase the data points available for investigating reproducibility, and to begin to identify properties of evaluations that are associated with better reproducibility.

We start in Section 2 by describing the organisation and structure of the shared task, followed by an overview of the participating teams (Section 3). Next, we present high-level degree-of-reproducibility results for each reproduction study, and in the case of the more complex studies, also for subsets of results (Section 4). We look at the properties of the ReproGen evaluation studies in standardised terms as facilitated by the HEDS sheets completed by participants, and explore if any properties appear to have an effect on degree of reproducibility (Section 5). We conclude with some discussion (Section 6) and a look to future work (Section 7).

## 2 ReproGen 2022

Like its predecessor, ReproGen 2022[2] had two tracks, one a shared task in which teams try to reproduce the same previous evaluation results, the other an 'unshared task' in which teams attempt to reproduce their own previous evaluation results:

A ***Main Reproducibility Track:*** For a shared set of selected evaluation studies, participants repeat one or more studies, and attempt to reproduce the results, using published information plus additional information and resources provided by the authors, and making commonsense assumptions where information is still incomplete.

B ***RYO Track:*** *R*eproduce *Y*our *O*wn previous evaluation results, and report what happened. Unshared task.

For the main track (A above), we used the same papers as in ReproGen 2021, with the addition of one paper (Nisioi et al., 2017) previously used

[1] See our systematic review of reproducibility research in NLP carried out in part as background research for ReproGen (Belz et al., 2021).

[2] All information and resources relating to ReproGen are available at https://reprogen.github.io/.

| Track | Team | Original paper | Reproduction paper | Metrics |
|---|---|---|---|---|
| A | Tilburg University | Santhanam and Shaikh (2019) | Braggaar et al. (2022) | automatic/human |
| | ADAPT Centre @ DCU | Nisioi et al. (2017) | Popović et al. (2022) | human |
| | University of Illinois at Chicago | Nisioi et al. (2017) | Arvan et al. (2022) | automatic |
| B | University of Aberdeen | Thomson and Reiter (2021) | Thomson and Reiter (2022) | human |
| | ADAPT, Charles Univ. Prague, Fed. Univ. of Minas Gerais | Dušek and Kasner (2020) | Huidrom et al. (2022) | human |

Table 1: Overview of ReproGen submissions (tracks, teams, original papers, reproduction reports and types of reproduced evaluation measures).

in the REPROLANG 2020 shared task (Branco et al., 2020), all with consent and confirmation of willingness to support from the authors:

1. van der Lee et al. (2017): *PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences*: 1 evaluation study; Dutch; 20 evaluators; 3 quality criteria; reproduction target: primary scores.
2. Dušek et al. (2018): *Findings of the E2E NLG Challenge*: 1 evaluation study; English; MTurk; 2 quality criteria; reproduction target: primary scores.
3. Qader et al. (2018): *Generation of Company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation*: 1 evaluation study; English; 19 evaluators; 4 quality criteria; reproduction target: primary scores.
4. Santhanam and Shaikh (2019): *Towards Best Experiment Design for Evaluating Dialogue System Output*: 4 evaluation studies differing in experimental design; English; 40 evaluators; 2 quality criteria; reproduction target: intraclass correlation between studies.
5. Nisioi et al. (2017): *Exploring Neural Text Simplification Models*: 1 evaluation study; English; 3 evaluators; 2 metric scores; 4 human-evaluated quality criteria; reproduction target: primary scores.

Authors of original papers in Track A were asked (i) to complete a HEDS datasheet[3] (Shimorina and Belz, 2022) for their paper, (ii) to make available all code and other resources needed for the study, and (iii) to be available to answer questions and provide other help during the ReproGen participation period. Authors of reproduction papers were also asked to complete a HEDS datasheet.

We issued a call for participation in one or both tracks. Six teams registered for ReproGen, of which five teams submitted reproduction studies (for an overview, see Table 1).

We made available broad guidelines[4] to participating teams about how to report reproduction results, and provided light-touch review with comments and feedback on papers.

## 3 Participants and Submissions

Five submissions were received by the deadline on June 6, 2022. One submission was from the Netherlands, one from the UK, one from the US, one from Ireland, and one was a collaboration between groups in Czechia, Brazil and Ireland. Three of the teams participated in Track A (Braggaar et al., 2022; Popović et al., 2022; Arvan et al., 2022); the other two in Track B (Thomson and Reiter, 2022; Huidrom et al., 2022).

Two of the submissions reported a reproduction study of Nisioi et al. (2017), one of Santhanam and Shaikh (2019), and two reproduced own earlier work. All of the evaluated systems produced outputs in English. Popović et al. and Arvan et al. reproduced the human and metric-based evaluations of Nisioi et al. (2017)'s simplification systems, respectively, with Arvan et al. additionally exploring variations in the system code. Braggaar et al. reproduced inter-rater agreement and consistency measures for human evaluations of a dialogue system involving different rating scales studied by Santhanam and Shaikh (2019). Thomson and Reiter looked at reproducing an evaluation by error annotation from their own work on data-to-text generation, using different evaluation data samples, and Huidrom et al. reproduced human evaluations of Dušek and Kasner (2020)'s semantic error detection system for data-to-text generation. An overview of all submissions is provided in Table 1, and the properties of participating systems and studies are discussed in more detail in Section 5.

[3] https://forms.gle/MgWiKVu7i5UHeMNQ9

[4] https://reprogen.github.io/2022/submission/

| Measurand(s) | Pearson's $r$ | Spearman's $\rho$ | mean % change | | mean CV* |
|---|---|---|---|---|---|
| | | | +/- | abs | |
| | | | | | |
| ***Original study = Nisioi et al. (2017); reproduction study = Arvan et al. (2022), Repro 1:*** | | | | | |
| All Scores (2 systems × 2 metrics) | 1 | 1 | 0 | 0 | 0 |
| | | | | | |
| ***Original study = Nisioi et al. (2017); reproduction study = Arvan et al. (2022), Repro 2:*** | | | | | |
| All Scores (2 systems × 2 metrics) | 1 | 0.8 | -1.02 | 3.30 | 3.34 |
| | | | | | |
| ***Original study = Nisioi et al. (2017); reproduction study = Arvan et al. (2022), Repro 3:*** | | | | | |
| All Scores (2 systems × 2 metrics) | 1 | 0.8 | 0.63 | 3.19 | 3.16 |
| | | | | | |
| ***Original study = Nisioi et al. (2017); reproduction study = Popović et al. (2022):*** | | | | | |
| All Scores (9 systems × 1 quality criterion) | $0.766^{**}$ | $0.787^{*}$ | 40.16 | 85.82 | 8.98 |
| | | | | | |
| ***Original study = Santhanam and Shaikh (2019); reproduction study = Braggaar et al. (2022):*** | | | | | |
| Likert (2 corr coeffs × 2 quality criteria × 1 scale) | $0.95^{*}$ | 0.81 | 25.37 | 25.37 | 21.88 |
| RME (2 corr coeffs × 2 quality criteria × 1 scale) | -0.57 | -0.54 | -6.895 | 6.895 | 7.25 |
| BME (2 corr coeffs × 2 quality criteria × 1 scale) | 0 | -0.07 | 8.55 | 8.55 | 8.15 |
| BWS (2 corr coeffs × 2 quality criteria × 1 scale) | $0.99^{**}$ | 0.88 | 10.02 | 10.02 | 9.52 |
| Readability (2 corr coeffs × 1 quality criterion × 4 scales) | -0.08 | 0.13 | 10.28 | 15.54 | 14.1 |
| Coherence (2 corr coeffs × 1 quality criterion × 4 scales) | -0.16 | 0.1 | 8.25 | 9.88 | 9.1 |
| ICC-C (1 corr coeffs × 2 quality criterion × 4 scales) | 0.33 | 0.5 | 8.12 | 10.24 | 9.67 |
| ICC-A (1 corr coeffs × 2 quality criterion × 4 scales) | -0.27 | -0.22 | 10.41 | 15.18 | 13.73 |
| All Scores (2 corr coeffs × 2 quality criteria × 4 scales) | 0.01 | 0.16 | 9.26 | 12.71 | 11.699 |
| | | | | | |
| ***Original study = Dušek and Kasner (2020); reproduction study = Huidrom et al. (2022), Repro 1:*** | | | | | |
| E2E (9 label counts × 1 system × 1 dataset) | $0.98^{**}$ | $0.91^{**}$ | 1.15 | 18.9 | 19.62 |
| WebNLG (8 label counts × 1 system × 1 dataset) | $0.8^{**}$ | $0.76^{*}$ | 41.46 | 70.12 | 50.89 |
| All Scores (8/9 label counts × 1 system × 2 datasets) | $0.81^{**}$ | $0.87^{**}$ | 20.12 | 43.00 | 34.34 |
| | | | | | |
| ***Original study = Dušek and Kasner (2020); reproduction study = Huidrom et al. (2022), Repro 2:*** | | | | | |
| E2E (9 label counts × 1 system × 1 dataset) | $0.87^{**}$ | $0.8^{*}$ | 18.57 | 40.45 | 32.32 |
| WebNLG (8 label counts × 1 system × 1 dataset) | $0.82^{**}$ | 0.54 | 18.97 | 58.17 | 46.86 |
| All Scores (8/9 label counts × 1 system × 2 datasets) | $0.84^{**}$ | $0.66^{**}$ | 18.76 | 48.79 | 39.16 |
| | | | | | |
| ***Original study = Thomson and Reiter (2021); reproduction study = Thomson and Reiter (2022), Repro 1:*** | | | | | |
| Cond-copy (6 label counts × 1 system) | 0.995 | 0.98 | 31.14 | 46.64 | 33.297 |
| Doc-plan (6 label counts × 1 system) | 0.91 | 0.90 | -7.92 | 16.50 | 48.88 |
| Hier-enc (6 label counts × 1 system) | 0.85 | 0.70 | 70.67 | 109.9 | 76.07 |
| All Scores (6 label counts × 3 systems) | 0.89 | 0.88 | 33.6 | 60.10 | 52.75 |
| | | | | | |
| ***Original study = Thomson and Reiter (2021); reproduction study = Thomson and Reiter (2022), Repro 2:*** | | | | | |
| Cond-copy (6 label counts × 1 system) | $0.99^{**}$ | $0.94^{*}$ | 31.79 | 57.37 | 46.73 |
| Doc-plan (6 label counts × 1 system) | $0.92^{**}$ | 0.82 | -24.35 | 29.18 | 68.57 |
| Hier-enc (6 label counts × 1 system) | $0.83^{*}$ | 0.72 | 73.86 | 136.64 | 88.70 |
| All Scores (6 label counts × 3 system) | $0.896^{**}$ | $0.84^{**}$ | 30.12 | 77.06 | 68.00 |

Table 2: Pearson's and Spearman's correlation coefficients, mean percentage change, and mean coefficients of variation (CV*), for the ReproGen'22 reproduction studies. For the correlation coefficients, $**$ = statistically significant at $\alpha = .01$, $*$ = statistically significant at $\alpha = .05$.

## 4 Results: Degree of Reproducibility

Table 2 shows summarising results for all submissions, or rather for every reproduction in every submission, i.e. nine original/reproduction study pairs, in terms of Pearson's r, Spearman's rho, mean percentage in/decrease, mean absolute percentage in/decrease, and the de-biased coefficient of variation, CV* (last column), following Belz (2022)'s Quantified Reproducibility Assessment (QRA) approach. The coefficient of variation (CV) is a

standard measure of precision used in metrological studies to quantify reproducibility of measurements. Unlike mean and standard deviation, CV is not in the unit of the measurements, and captures the amount of variation there is in a set of $n$ scores in a general way, providing a quantification of precision (degree of reproducibility) that is comparable across studies (Ahmed, 1995, p. 57). Note that all evaluation scales need to be shifted to start at zero, to ensure fair comparison across evaluations, because both percentage change and CV in general underestimate variation for scales with a lower end greater than 0. Rather than standard CV, QRA uses CV*, a de-biased version of CV (Belz, 2022), because sample size (number of repeat measures) tends to be very small in NLP.[5]

For the simpler reproductions in Table 2, where there were one or more systems and one or more conventional evaluation measures and the reproduction target was the overall scores in terms of the measure(s), Table 2 reports a single CV* figure in the last column, namely mean CV* over all systems and measures. For example, the fourth study in the table, Popović et al. (2022)'s reproduction of Nisioi et al. (2017), has an overall mean CV* of 8.98, computed from 9 individual CV* figures (9 systems $\times$ 1 quality criterion).

For the five remaining studies, we also show mean CV* for constituent subsets of individual CV* figures, grouped by rating scale, quality criterion and correlation coefficient for Braggaar et al. (2022)'s reproductions, by dataset for Huidrom et al. (2022)'s reproductions, and by system for Thomson and Reiter (2022)'s reproductions.

Columns 2 and 3 in Table 2 show Pearson's r and Spearman's rho, respectively, for the corresponding (sub)sets of original/reproduction score pairs, while Columns 4 and 5 show average percentage in/decrease from original to reproduction score pairs for each of the same (sub)sets.

We have ordered the studies by study-level mean CV* (lowest, i.e. best, first). Study-level mean CV* ranges from the perfectly reproduced metric scores in Arvan et al. (2022)'s first reproduction, to the particularly high CV* of Thomson and Reiter (2022)'s second reproduction of an error annotation. In the case of the former, the authors managed to obtain the exact same SARI and BLEU scores, by running the scripts for these metrics provided by the original authors on the system outputs also provided by the original authors. Thomson and Reiter (2022)'s reproductions involve error-type labelling of system outputs which appear to be a particularly difficult to reproduce form of evaluation: this was the reproduction target in the four studies in the lower half of Table 2 which have substantially higher (>34) resulting overall mean CV* than the other studies (<12).

Interpreting the mean CV* figures for subsets of results for Braggaar et al. (2022)'s reproduction is not simple. The original authors collected evaluations of a set of dialogue turns in context for 2 quality criteria (Readability and Coherence), repeated this for 4 different rating scales, and computed two measures of inter-rater similarity for each rating scale. The two measures of inter-rater similarity were the consistency intraclass correlation (ICC-C) and the agreement intraclass correlation (ICC-A). The mean CV* figures for Braggaar et al. (2022)'s reproduction in Table 2 thus measure the similarity between the ICC scores (automatically computed on the human ratings) in the original study and the ICC scores in the reproduction study, with the ICC scores themselves computed for each set of ratings (where each set corresponds to one of the scales combined with one of the quality criteria).

Under these circumstances, CV* expresses how reproducible (stable) the inter-rater consistency/agreement is from one experiment to a repetition of it, in other words whether inter-rater consistency/agreement is similarly high, or similarly low, across multiple repeats of the same evaluation. Because Braggaar et al. (2022) repeated the evaluations for four different rating instruments, the mean CV* figures can tell us whether this differs for different rating instruments (as well as for different evaluation criteria and inter-rater consistency/agreement measures). The answer is that it does differ substantially for different rating scales, is equally low for both evaluation criteria, and does differ for the two inter-rater measures.

Taking a slightly closer look, the inter-rater measures (ICCs) for the Likert scale have remarkably higher (worse) mean CV* than the other three scales, while nevertheless achieving strong Pearson's and Spearman's between individual ICC scores in the original and reproduction studies. While the ICCs for the other three scales have similarly good CV*, only the BWS scale also has strong Pearson's and Spearman's, with BME having no

[5]For full details of, and rationale for, using CV*, even for sets of just two scores, see Belz et al. (2022); Belz (2022).

correlation and RME having medium-strength *negative* correlation. This shows that CV* and Pearson's and Spearman's correlation coefficients provide complementary information in assessments of the similarity of original vs. reproduction scores. Looking at these in combination, it would seem that the BWS scale (best-to-worst ranking) achieves the most similar levels of inter-rater agreement and consistency across repeat studies.

From the results for Huidrom et al. (2022)'s reproductions, we can see that the error annotations produced for outputs for WebNLG data have worse CV* figures than for E2E (the difference is not just the data but also a subset of the error categories which are tailored to the data). Here, better CV* is aligned with better correlations.

Finally, the results for Thomson and Reiter (2022)'s reproductions show that the hierarchical encoder based data-to-text system produced outputs for which both mean CV* and correlations were worse on average than for the other two systems. However, this latter observation ought to be read with the proviso that each reproduction used a different sample from the three systems.

## 5 Comparison of Properties of Original vs. Reproduction Studies

Overall, all teams tried to follow the original studies as closely as possible (see also Discussion section below), but cohorts of human evaluators involved were different across all pairs of original and reproduction studies, except for the two reproductions by Thomson and Reiter (2022), and one of the two by Huidrom et al. (2022).

In this section, we summarise differences in each pair of studies and highlight the possible factors that might have led to different results in reproduction results. In the case of Track A contributions, our notes are based on the HEDS datasheets completed by both the original study authors and the shared task participants. For Track B, we describe differences as reported by the authors themselves in their original and reproduction reports, also consulting the HEDS sheets completed by them. See also Table 3 which lists some of the more fine-grained information for each study from the HEDS sheets.

### 5.1 Track A

Popović et al. (2022) reproduced the human evaluation reported by Nisioi et al. (2017), and point out the following differences that might have influenced the reproduction: evaluator background (native language, profession, experience with text simplification evaluation), evaluator assignments to texts, and experimental setup (e.g. whether evaluators were allowed to ask questions about guidelines), all of which were not reported for the original study, and not obtainable from the original authors.

Arvan et al. (2022) also reproduced Nisioi et al. (2017)'s work, but just the metric scores. They focused on exploring different ways of obtaining the outputs to be evaluated (having discovered several substantial issues with the original code): (a) using the same outputs, (b) regenerating outputs with the same code, and (c) regenerating outputs with corrected code. They found an "extreme level of resilience [to such differences that] is, in fact, quite alarming," which is reflected in the low mean CV* figures which as it happens also reflect variation from different versions of SacreBLEU.

Braggaar et al. (2022)'s reproduction of Santhanam and Shaikh (2019) used crowdsourced human evaluation like the original study, but on a different platform: Qualtrics and Prolific in the reproduction study, and MTurk in the original. Due to platform feature restrictions, questionnaire layouts were not exactly the same across the two studies. As for the inter-rater measures, Braggaar et al. wrote their own code to compute ICC scores, since it was not provided by the original authors.

### 5.2 Track B

Huidrom et al. (2022) carried out two reproduction studies of Dušek and Kasner (2020): the first one with the same two evaluators and the second one with two new evaluators. The main difference between the original and reproduction studies lies in error annotation guidelines and output assignments to evaluators. While the original study did not formalise the annotation guidelines and performed evaluation based on common understanding developed between the two evaluators, for the reproduction studies, instructions for applying the error annotation scheme were created and used. The original study also did not record which texts were evaluated by which annotator, so the reproduction studies randomly assigned annotators to evaluated texts.

The main difference between the two reproductions and the original work addressed by Thomson and Reiter (2022) was the use of different samples

of outputs albeit from the same larger test set. This did result in substantial differences between results, as we shall see below.

### 5.3 Study properties and reproducibility

Table 3 provides an overview of the five ReproGen'22 submissions in terms of the quality criteria assessed in the evaluations and the properties of the evaluation design. The first column identifies the study and criteria, the last column shows the corresponding mean study-level and mean criterion-level CV*. The remaining columns show seven properties of each study/criterion, as per the HEDS datasheets; column headings identify HEDS question number (for explanation of each see table caption). The lower half of the table shows the corresponding overview of study/criterion properties from ReproGen'21, for ease of comparison.

In the ReproGen'22 studies, annotation-based evaluation (4.3.8=Anno) is clearly associated with lower reproducibility. Evaluations which involve assessment of content alone (4.1.2=Cont) also tend to have worse reproducibility. Assessing evaluation items relative to a system input (4.1.3=RtI) is also associated with lower reproducibility for the bottom three studies (where comparison of outputs to inputs is far more complex than a straightforward is-it-simpler decision as in *Nisioi et al/Popovic et al*). Finally, correctness assessment (4.1.1=Corr) is also associated with lower reproducibility. For those of these properties that were present in ReproGen'21, the tendencies are the same.

## 6 Discussion

In metrological terms, a *repeatability* assessment keeps all conditions under which a measurement was taken the same, whereas a *reproducibility* assessment varies some of them. Strictly speaking, only the first reproduction by Arvan et al. (2022) can be considered a repeatability assessment, as it keeps all conditions exactly the same. All other ReproGen'22 reproductions were human evaluations, and for these, conditions can only be the same if the same evaluators are used again. One of the studies (the first reproduction by Huidrom et al. (2022)) did use the same evaluators, but instructions were written down and used for the first time instead of evaluators conferring.

Nevertheless, all studies tried to keep things as much as possible the same. One study which looked at automatic metrics (only) (Arvan et al., 2022) went beyond reusing system outputs provided by original authors, (a) regenerating outputs with unchanged author-provided code, and (b) regenerating outputs with a retrained system, including with a substantial correction to the code. Interestingly, evaluation results were very similar in all versions where outputs were regenerated, including switching word2vec embeddings on/off.

For the studies looking at human evaluations, new cohorts of evaluators were rarely able to achieve low CV* scores, generally only in very simple assessments. Pearson and Spearman correlations were generally better, with some exceptions where comparison was between inter-rater similarity measures, rather than evaluation scores (*Santhanam & Shaikh/Braggaar et al*).

We saw that correlation coefficients and mean CV* often but not always give the same indication of similarity between a set of original and reproduction scores. For example, the results in Table 2 for Braggaar et al. (2022)'s reproduction of Santhanam and Shaikh (2019) show that for Likert we have high correlation but poor CV*, for RME and BME, correlation is inverse or absent, but CV* is good, and for BWS both are good. For all other studies, better CV* always means better correlations.

This year we had a few (new) firsts at ReproGen, in addition to automatic metrics being reproduced for the first time: e.g. Thomson and Reiter (2022) investigated the effect of swapping out the data sample (from the same superset), while keeping all other conditions the same including annotators. As the sample size is fairly small, and differed in size between original study and the two reproductions, it's perhaps not surprising that error label counts varied substantially between studies.

Some of the ReproGen'22 participants' reports mention less than ideal support from original authors during reproductions, despite the fact that all original authors had agreed to support and help with ReproGen'22 reproductions. Of course, such help is essential to testing reproducibility, and in future shared tasks, we will consider the option of obtaining more of the resources and information prior to the start of the shared task.

## 7 Conclusions

We first proposed the ReproGen shared task at Generation Challenges 2020[6] (Belz et al., 2020) and, taking into account feedback received, developed it

[6] INLG'20, Dublin.

| ReproGen 2022 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Studies,* **measurands** | **3.1.1** | **3.2.1** | **4.3.4** | **4.3.8** | **4.1.1** | **4.1.2** | **4.1.3** | **scores /item** | **mean CV*** |
| *Nisioi et al / Arvan et al 1–3* | | | | | | | EFoR | | 2.17 |
| **SARI** | ∼50 | NA/NA | [0..1] | DQE | Good | Form | +RtI | NA | 2.34 |
| **BLEU** | ∼50 | NA/NA | [0..1] | DQE | Good | Form | EFoR | NA | 1.99 |
| *Nisioi et al / Popovic et al,* **Simplicity** | 70 | 3/3 | -2, -1, 0, 1, 2 | DQE | Feature | Both | RtI | 2 | 8.98 |
| *Santhanam & Shaikh / Braggaar et al* | | | | | | | | | 11.7 |
| **ICC for Readability** | | | | | | | | | 14.1 |
| Likert scale | 50 | 160/163 | 1–6 | DQE | Good | Both | iiOR | 1 | 28.19 |
| Magnitude est. (stdval=100) | 50 | 160/163 | 100 | DQE | Good | Both | iiOR | 1 | 11.18 |
| Magnitude est. (stdval=var) | 50 | 160/163 | 100 | DQE | Good | Both | iiOR | 1 | 6.93 |
| Best-to-worst ranking | 50 | 160/163 | 4! rankings | RQE | Good | Both | iiOR | 1 | 10.1 |
| **ICC for Coherence** | | | | | | | | | 9.3 |
| Likert scale | 50 | 160/163 | 1–6 | DQE | Good | Cont | iiOR | 1 | 15.58 |
| Magnitude est. (stdval=100) | 50 | 160/163 | 100 | DQE | Good | Cont | iiOR | 1 | 3.31 |
| Magnitude est. (stdval=var) | 50 | 160/163 | 100 | DQE | Good | Cont | iiOR | 1 | 9.38 |
| Best-to-worst ranking | 50 | 160/163 | 4! rankings | RQE | Good | Cont | iiOR | 1 | 8.93 |
| *Dusek & Kasner / Huidrom et al 1&2* | | | | | | | | | 36.75 |
| **Label counts from correctness annotations** | 200 | 2/2 | 3 labels | Anno | Corr | Cont | RtI | 1 | 18.11 |
| **Label counts from error type annotations** | 200 | 2/2 | 6/5 labels | Anno | Corr | Cont | RtI | 1 | 46.92 |
| *Thomson & Reiter / Thomson & Reiter 1 & 2*, **Label counts from error type annotations** | 13, 10 | 3/3 | 6 labels | Anno | Corr | Cont | RtI | 3 | 68 |
| ReproGen 2021 | | | | | | | | | |
| Lee et al./Mille et al. | | | | | | | | | 11.89 |
| Stance ID Acc | 10 | 20/20 | stance A, stance B | output classif | Feature | Both | EFoR | 20 | 6.11 |
| Clarity S3 ('Understandability') | 20 | 20/20 | 1–7 | DQE | Good | Both | iiOR | 20 | 12.03 |
| Clarity S4 ('Clarity') | 20 | 20/20 | 1–7 | DQE | Good | Both | iiOR | 20 | 14.61 |
| Fluency S1 ('Grammaticality') | 20 | 20/20 | 1–7 | DQE | Corr | Form | iiOR | 20 | 18.3 |
| Fluency S2 ('Readability') | 20 | 20/20 | 1–7 | DQE | Good | Both | iiOR | 20 | 13.71 |
| Popović/Popović & Belz | | | | | | | | | 29.22 |
| Comprehension Minor | 557, | 7/7 | 2 labels | Anno | Good | Both | iiOR | 2 | 22.14 |
| Comprehension Major | 279, | 7/7 | | Anno | Good | Both | iiOR | 2 | 38.23 |
| Adequacy Minor | 467 | 7/7 | 3 labels | Anno | Corr | Cont | RtI | 2 | 17.83 |
| Adequacy Major | | 7/7 | | Anno | Corr | Cont | RtI | 2 | 38.67 |
| Qader et al./Richter et al. | | | | | | | | | 22.16 |
| Information Coverage | 30 | 19/19 | 1–5 | DQE | Corr | Cont | RtI | 1 | 34.04 |
| Information Non-redundancy | 30 | 19/19 | 1–5 | DQE | Good | Cont | iiOR | 1 | 19.11 |
| Semantic Adequacy | 30 | 19/19 | 1–5 | DQE | Corr | Cont | iiOR | 1 | 20.4 |
| Grammatical Correctness | 30 | 19/19 | 1–5 | DQE | Corr | Form | iiOR | 1 | 15.09 |
| Mahamood et al./Mahamood, Binary Preference Strength | 2† | 25‡/11 | -3..+3 | RQE | Good | Both | EFoR | 25/11 | 72.34 |

Table 3: Summary of some properties from HEDS datasheets provided by ReproGen participants (in some cases corrected by organisers. 3.1.1 = number of items assessed per system; 3.2.1 = number of evaluators in original/reproduction experiment; 4.3.4 = List/range of possible responses; 4.3.8 = Form of response elicitation (DQE: direct quality estimation, RQE: relative quality estimation, Anno: evaluation through annotation); 4.1.1 = Correctness/Goodness/Features; 4.1.2 = Form/Content/Both; 4.1.3 = each output assessed in its own right (iiOR) / relative to inputs (RtI) / relative to external reference (EFoR); scores/item = number of evaluators who evaluate each evaluation item; mean CV*. † considering texts with and without hedges to be the two systems being compared. ‡ subset of 32 evaluators from original studies: 14 native + 11 fluent speakers.

into the two iterations of ReproGen, 2021 and 2022, the latter reported in the present paper. ReproGen was intended as a testbed for an NLP-wide shared task on reproduction, and in 2023 we intend to run

an expanded version, the ReproHum Shared Task on Reprodubility of Evaluation Results in NLP, initially for just human evaluations.

We have gained some important insights from ReproGen, in particular with regard to what kind of properties of evaluations tend to increase or decrease degree of reproducibility. Perhaps not surprisingly, it is very clear that the lower the cognitive load on evaluators while making individual assessments, the better reproducibility.

In a research culture that prizes leaderboard success, it was always going to be difficult to incentivise people to carry out tasks that are basically just good scientific hygiene, but we hope we have made a contribution to raising awareness of the importance of having reproducible evaluations, and of testing our methods for reproducibility. After all, how else are we going to know for sure that one approach is better than another.

## Acknowledgments

We thank the authors of the five original papers that were up for reproduction in Track A. And of course the authors of the reproduction papers, without whom there would be no shared task.

Our work was carried out as part of the ReproHum project on Investigating Reproducibility of Human Evaluations in Natural Language Processing, funded by EPSRC (UK) under grant number EP/V05645X/1.

Popović's work is directly funded by the ADAPT SFI Centre for Digital Media Technology which is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106. Both Popović and Belz also benefit in other ways from being members of ADAPT.

## References

S. E. Ahmed. 1995. A pooling methodology for coefficient of variation. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 57–75.

Mohammad Arvan, Luís Pina, and Natalie Parde. 2022. Reproducibility of exploring neural text simplification models: A review. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, Waterville, Maine, USA. Association for Computational Linguistics.

Anya Belz. 2022. A metrological perspective on reproducibility in nlp. *Computational Linguistics*, 48.

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.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Anya Belz, Maja Popović, and Simon Mille. 2022. Quantified reproducibility assessment of nlp results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16–28.

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*, Waterville, Maine, USA. Association for Computational Linguistics.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with REPROLANG2020. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5539–5545, Marseille, France. European Language Resources Association.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the E2E NLG challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 322–328, Tilburg University, The Netherlands. Association for Computational Linguistics.

Rudali Huidrom, Ondřej Dušek, Zdeněk Kasner, Thiago Castro Ferreira, and Anya Belz. 2022. Two reproductions of a human-assessed comparative evaluation of a semantic error detection system. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, Waterville, Maine, USA. Association for Computational Linguistics.

Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2017. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104, Santiago de Compostela, Spain. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Maja Popović, Sheila Castilho, Rudali Huidrom, and Anya Belz. 2022. Reproducing a manual evaluation of simplicity in text simplification system outputs. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, Waterville, Maine, USA. Association for Computational Linguistics.

Raheel Qader, Khoder Jneid, François Portet, and Cyril Labbé. 2018. Generation of company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 254–263, Tilburg University, The Netherlands. 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.

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.

Craig Thomson and Ehud Reiter. 2021. Generation challenges: Results of the accuracy evaluation shared task. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 240–248, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Craig Thomson and Ehud Reiter. 2022. The accuracy evaluation shared task as a retrospective reproduction study. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, Waterville, Maine, USA. Association for Computational Linguistics.