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

Anya Belz
ADAPT Research Centre, DCU, Ireland
anya.belz@adaptcentre.ie

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

Shubham Agarwal
Heriot Watt University, UK
sa201@hw.ac.uk

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

Abstract

The NLP field has recently seen a substantial increase in work related to reproducibility of results, and more generally in recognition of the importance of having shared definitions and practices relating to evaluation. Much of the work on reproducibility has so far focused on metric scores, with reproducibility of human evaluation results receiving far less attention. As part of a research programme designed to develop theory and practice of reproducibility assessment in NLP, we organised the first shared task on reproducibility of human evaluations, ReproGen 2021. This paper describes the shared task in detail, summarises results from each of the reproduction studies submitted, and provides further comparative analysis of the results. Out of nine initial team registrations, we received submissions from four teams. Meta-analysis of the four reproduction studies revealed varying degrees of reproducibility, and allowed very tentative first conclusions about what types of evaluation tend to have better reproducibility.

1 Introduction

There has been growing interest in reproducibility across Natural Language Processing (NLP) over recent years.[1] However, work has mostly focused on determining what information and resources need to be shared to enable others to obtain the same metric results. The reproducibility of human evaluation has received far less attention and currently very little is known about how reproducible, hence trustworthy, the human evaluations we routinely apply in NLP really are. This is of particular concern in Natural Language Generation (NLG) where human evaluations have always played a central role (Reiter, 2018; Novikova et al., 2017).

The last few years have seen a growth in publications, projects, workshops, shared tasks and other initiatives on the topic of reproducibility. For example, NeurIPS'19 introduced the ML Reproducibility checklist for submitted papers (Pineau et al., 2020) which was also adopted by EMNLP'20 and AAAI'21. The Reproducibility Challenge has been running since 2018, initially in conjunction with ICLR then NeurIPS (Sinha et al., 2020). The Challenge is focused on ML results and metric scores, and is organised as a 'live' challenge, where participants pick one of the accepted papers, and try to reproduce its ML results (Sinha et al., 2020).

The REPROLANG'20 shared task (Branco et al., 2020) asked participants to reproduce results from 11 papers in different areas of NLP. While in the case of ten of the papers, the results up for reproduction were automatic scores, in one case (Nisioi et al., 2017) they included human evaluation scores.[2] In their reproduction study of this work, Cooper and Shardlow (2020) reannotated original system outputs using their own annotators, in order to be able to compare annotation results. Their results suggested a drop in both quality metrics of close to 15%.

Apart from the above reproduction study involving text simplification carried out within REPROLANG, there appears to have been just one other paper reporting reproduction studies of human evaluation results in NLG (Belz and Kow, 2011) which re-ran two evaluation experiments

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

[2] Task D.1: Text simplification: http://wordpress.let.vupr.nl/lrec-reproduction/

Proceedings of the 14th International Conference on Natural Language Generation (INLG), pages 249–258,
Aberdeen, Scotland, UK, 20-24 September 2021.

with the same evaluator cohorts, one in data-to-text generation, the other in visual referring expression generation. Here, strong correlations between annotator scores were found for two quality criteria for each task, 0.87 Pearson's in one case, >0.94 in the other three.

With the ReproGen shared task, our aim was to add to this currently small body of literature, in order to shed more light on how reproducible current human evaluation methods are, and what we may need to change in how we design and carry out human evaluations in order to improve reproducibility. In Section 2 we start by describing the organisation and structure of the shared task. Next we provide an overview of the participating teams (Section 3) and look at the properties of submitted systems (Section 4). We compare and analyse the results from the submitted systems in detail (Section 5), before we conclude with some discussion (Section 6) and tentative conclusions (Section 7).

2 Organisation of Shared Task

ReproGen'21[3] had two tracks, one a shared task in which teams try to reproduce the same prior human evaluation results, the other an 'unshared task' in which teams attempt to reproduce their own prior human evaluation results:

A ***Main Reproducibility Track:*** For a shared set of selected human 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 common-sense assumptions where information is still incomplete.

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

For the main track (A above), we issued a call for proposals of papers, asking people to propose papers via an online form.[4] This yielded seven proposed papers, from which we selected four on the grounds of suitability for reproduction studies, diversity of languages and cost of reproduction. The selected papers and studies, with many thanks to the authors for supporting ReproGen, are:

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*: 3 evaluation studies differing in experimental design; English; 40 evaluators; 2 quality criteria; reproduction target: correlation scores between 3 studies.

Authors of original papers in Track A were asked (i) to complete a HEDS datasheet[5] (Shimorina and Belz, 2021) 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, inviting teams to participate in one or both tracks. Nine teams registered for ReproGen, with team members from five different countries, out of which four teams submitted reproduction studies. Details of the submitting teams can be found in the following section.

We made available broad guidelines[6] to participating teams about how to report reproduction results, and provided light-touch review with comments and feedback on papers.

3 Overview of Participants and Submissions

Four submissions were received by the deadline on August 15, 2021. Two of the submissions were from Germany, one from Ireland, and one was a collaboration between groups in Spain, Brazil and Ireland. Two of the teams participated in Track A

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

[4] https://forms.gle/J5ranvXqmfjPDbxLA

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

[6] https://reprogen.github.io/submissions/

Track	Team	Original paper	Reproduction paper
A	Technical University of Darmstadt (TUDA)	Qader et al. (2018)	Richter et al. (2021)
	UPF Barcelona, UF Minas Gerais, ADAPT Dublin	van der Lee et al. (2017)	Mille et al. (2021)
B	Trivago GmbH, Düsseldorf	Mahamood et al. (2007)	Mahamood (2021)
	ADAPT Dublin	Popović (2020)	Popović and Belz (2021)

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

(Mille et al., 2021; Richter et al., 2021), the other two in Track B (Mahamood, 2021; Popović and Belz, 2021). Three of the four teams are affiliated with universities, one with a commercial company.

Each of the submissions reported a reproduction study for a different paper. Two of the evaluated systems produced outputs in English, one in Croatian, and one in Dutch. While Mahamood (2021) and Mille et al. (2021) reproduced human evaluation of data-to-text systems, Popović and Belz (2021) evaluated Machine Translation (MT) systems and Richter et al. (2021) text-to-text and concept-to-text generation systems. An overview of all submissions is provided in Table 1, and the properties of participating systems and studies are discussed in more detail in the next section.

4 Comparison of Properties of Original vs. Reproduction Studies

Overall, all teams tried to follow the original studies as closely as possible. All of the reproduction studies evaluated the same texts as reported in the original experiments, with the same criteria and measurement methods. Three of the four submissions used the same number of evaluators. Cohorts of human evaluators involved were different across all pairs of original and reproduction studies.

Below we summarise differences in each pair of studies and highlight the possible factors that might have affected 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.

4.1 Track A

Mille et al. (2021) reproduced van der Lee et al. (2017), the main differences being recruitment process and means of response collection. The original study recruited people on campus where they filled paper forms in one sitting, whereas the reproduction study used online surveys, where there was no control for timing, and people were recruited via personal contacts, i.e. they also included people known to the authors. The online form the authors used was designed to resemble the original paper form as much as possible. In addition, the reproduction study carried out some quality checks after the survey completion and replaced one entry from one participant, while the original experiment did not have any quality assurance methods (and consequently had some missing values).

Richter et al. (2021) reproduced Qader et al. (2018). Similar to the previous reproduction study of Mille et al. (2021), the main differences lie in survey design, and participant recruitment and background. While the original study used a specific web-based interface, the reproduction study built a Google form. That led to some differences in the interface, e.g. using checkboxes instead of a slider in the original evaluation. As regards human participants, the original evaluation was circulated among the authors' colleagues in their research lab; in contrast, the reproduction was carried out with friends and acquaintances. Both studies assessed English text in non-English-speaking countries; there was no formal assessment of the level of English among participants. Finally, manual quality checking was present in both studies after the evaluation experiment (for details, see the two papers); this involved subjective judgements and is hard to repeat across two studies.

4.2 Track B

Mahamood (2021) reproduced Mahamood et al. (2007). The original study used paper forms, while the reproduction used an online form. Evaluators were Master students in the original; the reproduction study instead used work colleagues. Another difference consists in the number of evaluators involved. There were 25 participants in the part of the original study that was reproduced; in contrast, the

Measurand(s)	Pearson's r	Spearman's ρ	mean % change		mean CV*
			+/-	abs	
Original study = van der Lee et al. (2017); reproduction study = Mille et al. (2021):					
All scores (1 system × 3 measures)	0.999**	1	10.19	10.19	11.891
Original study = Mahamood et al. (2007); reproduction study = Mahamood (2021):					
All scores (2 scenarios × 2 evaluator cohorts)	0.085	0.4	-24.14	60.16	72.343
Original study = Popović (2020); reproduction study = Popović and Belz (2021):					
Comprehension Minor, % words with errors (3 systems)	0.666	0.993	26.033	26.033	22.143
Comprehension Major, % words with errors (3 systems)	0.988*	0.973	47.953	47.953	38.227
Adequacy Minor, % words with errors (3 systems)	0.362	0.277	0.350	17.210	17.830
Adequacy Major, % words with errors (3 systems)	0.9986**	0.9997	48.443	48.443	38.667
All Scores (3 systems × 4 measures)	0.691**	0.818**	30.695	34.910	29.217
Original study = Qader et al. (2018); reproduction study = Richter et al. (2021):					
Mean Information Coverage (7 systems)	0.567	0.3397	36.826	42.840	34.044
Mean Non-redundancy of Information (7 systems)	0.328	0.524	1.899	19.153	19.108
Mean Semantic Adequacy (7 systems)	0.514	0.378	-2.979	19.201	20.396
Mean Grammatical Correctness (7 systems)	0.322	0.136	4.600	16.003	15.089
All Scores (7 systems × 4 measures)	0.679**	0.343	10.086	24.299	22.159

Table 2: Pearson's and Spearman's correlation coefficients, mean percentage change, and mean coefficients of variation (CV*), for the ReproGen'21 reproduction studies. ∗∗ = statistically significant at $\alpha = .01$, ∗ = at $\alpha = .05$.

reproduction study had 11 evaluators. Furthermore, the ratios between native and fluent English speakers were not the same: 14 and 11 in the original vs. 5 and 6 in the reproduction. Such distinctions may impact the reproduction results, since the experiment examines the effect of hedges on native versus fluent English speakers.

Popović and Belz (2021) carried out a reproduction study of Popović (2020). The reproduction study followed the original closely, with the main difference in participant background. While students and researchers in computational linguistics with different levels of MT experience took part in the original study, the reproduction study involved translation students with roughly the same level of MT experience.

5 Comparing Reproducibility in the ReproGen Studies

Table 4 shows results from all submissions, in terms of the individual pairs of scores reported in original and reproduction paper (columns 2 and 3), the percentage increase or decrease from original to reproduction score (column 4), and the de-biased coefficient of variation, CV* (last column), following Belz (2021). 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 we have shifted all evaluation scales 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, we use CV*, a de-biased version of CV, Belz (2021), because sample size (number of repeat measures) tends to be very small in NLP.[7]

CV* in Table 4 ranges from 6.107 to 16.372 for Mille et al. (2021)'s reproduction study; from 52.806 to 101.894 for Mahamood (2021); from 4.86 to 47.17 for Popović and Belz (2021); and from 0 to 66.467 for Richter et al. (2021). Percentage change gives a similar picture, as the two measures generally give similar results for sample size 2 (Pearson's correlation for absolute percentage change and CV* is 0.89 over all scores in Table 4).

Looking at the above CV* ranges for each reproduction study, a first indication of a ranking emerges for the four study pairs in terms of degree of reproducibility, with (1) Lee et al./Mille et al. having the highest degree of reproducibility, followed by (2) Popović/Popović & Belz, (3) Qader et al./Richter et al., and finally (4) Mahamood et al./Mahamood.

Table 2 provides higher-level results, where in each case multiple score pairs are analysed jointly,

[7]For full details of, and rationale for, using CV*, even for sets of just two scores, see Belz (2021).

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*
Lee et al./Mille et al.									11.891
Stance ID Acc	10	20/20	stance A, stance B	output classif	Feature	Both	EFoR	20	6.107
Clarity S3 ('Understandability')	20	20/20	1–7	DQE	Good	Both	iiOR	20	12.031
Clarity S4 ('Clarity')	20	20/20	1–7	DQE	Good	Both	iiOR	20	14.605
Fluency S1 ('Grammaticality')	20	20/20	1–7	DQE	Corr	Form	iiOR	20	18.303
Fluency S2 ('Readability')	20	20/20	1–7	DQE	Good	Both	iiOR	20	13.711
Popović/Popović & Belz									29.217
Comprehension Minor	557,	7/7	2 labels	Anno	Good	Both	iiOR	2	22.143
Comprehension Major	279,	7/7	2 labels	Anno	Good	Both	iiOR	2	38.227
Adequacy Minor	467	7/7	3 labels	Anno	Corr	Cont	RtI	2	17.830
Adequacy Major		7/7	3 labels	Anno	Corr	Cont	RtI	2	38.667
Qader et al./Richter et al.									22.159
Information Coverage	30	19/19	1–5	DQE	Corr	Cont	RtI	1	34.044
Information Non-redundancy	30	19/19	1–5	DQE	Good	Cont	iiOR	1	19.108
Semantic Adequacy	30	19/19	1–5	DQE	Corr	Cont	iiOR	1	20.396
Grammatical Correctness	30	19/19	1–5	DQE	Corr	Form	iiOR	1	15.089
Mahamood et al./Mahamood, Binary Preference Strength	2†	25‡/11	-3..+3	RQE	Good	Both	EFoR	25/11	72.343

Table 3: Summary of some properties from HEDS datasheets provided by ReproGen participants. 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.

in terms of Pearson's and Spearman's correlation coefficients (columns 2 and 3), mean percentage change and mean *absolute* percentage change (columns 4 and 5), and mean CV* (last column). For example, for Lee et al./Mille et al., Pearson's r was 0.99 for the three scores in the original study compared with the corresponding scores from the reproduction study, both as shown in Table 4; Spearman's ρ was 1 (i.e. all ranks were the same); on average scores went up by 10.19%; the absolute percentage change was also 10.19% (because all changes were positive); and on average CV* was 11.89. Where a study compared multiple systems in absolute terms,[8] we show results per evaluation measure (e.g. Comprehension Minor), in addition to results for all scores.

In terms of the study-level scores ('All Scores' rows) in Table 2, a more mixed picture emerges compared to Table 4. In terms of both Pearson's and Spearman's, the ranking is the same in Table 2 and Table 4: (1) Lee et al./Mille et al., (2) Popović/Popović & Belz, (3) Qader et al./Richter et al., then (4) Mahamood et al./Mahamood. In contrast, the rankings for overall mean (absolute) percentage change and overall mean CV* are slightly different: (1) Lee et al./Mille et al., (2) Qader et al./Richter et al., (3) Popović/Popović & Belz, then (4) Mahamood et al./Mahamood.

In Table 3, we summarise some properties of our four pairs of studies, in terms of a subset of the properties from the HEDS datasheet (Shimorina and Belz, 2021) we asked participants to complete,[9] to attempt to identify possible relationships between study properties and degree of reproducibility. As discussed in the next section, such interpretations could be made with greater confidence if sample sizes were larger than 2, and we intend to add further studies in the future to enable more confident conclusions.

Something that's not easily captured in a table is the differences in cohorts of evaluators. For example, in Mahamood et al./Mahamood, evaluators in the original study were students, whereas non-students were used in the reproduction study; the former were a lot younger on average. In Lee et al./Mille et al., the original study used random people encountered in the university's science building, the reproduction study used present and

[8] Mahamood et al./Mahamood assess two systems, but in relative terms, yielding just one score.

[9] We corrected the information provided in a small number of cases by referring to the papers.

former staff and postgraduate students in computing science some of whom were known to the authors; here too the former were a lot younger on average. In Popović/Popović & Belz, the original evaluators were computational linguistics staff and students, the evaluators in the reproduction study were translation students. Finally, in Qader et al./Richter et al., the original evaluators were recruited from among people in the same lab (excluding the authors), whereas the reproduction study authors recruited people from their social environment. Broadly speaking, differences between evaluator cohorts would seem to be particularly pronounced in Qader et al./Richter et al. and Mahamood et al./Mahamood, and these two study pairs are also the least reproducible out of the four study pairs, according to all measures except *mean* absolute percentage change and *mean* CV*.

In Table 3, column 2 shows the number of items assessed per system (Question 3.1.1 in the HEDS datasheet); column 3 shows the number of evaluators in an experiment (Question 3.2.1 in HEDS); column 4 shows the list/range of possible responses (4.3.4); column 5 shows the form of response elicitation (4.3.8); column 6 shows whether the underlying quality criterion assesses the correctness, goodness, or a feature-type aspect of quality (4.1.1); column 7 shows whether the quality criterion assesses an output's form, content or both (4.1.2); column 8 shows whether the quality criterion assesses each output in its own right (iiOR), relative to input (RtI), or relative to an external frame of reference (EFoR) (4.1.3); column 9 ('scores/item') shows the number of scores collected per evaluation item; finally, the last column shows corresponding mean CV* values for ease of reference. For full details regarding HEDS questions and possible values, see Shimorina and Belz (2021).

In Lee et al./Mille et al., Clarity and Fluency are compound measures each derived from two separately assessed quality criteria, which map to the normalised quality criterion names shown in rows 4–7 in Table 3, following the taxonomy of normalised quality criteria proposed by Howcroft et al. (2020).

Looking at Table 3, it's hard to detect any specific patterns in study properties that might be predictive of CV*. There is perhaps some indication that the (normalised) Grammaticality criterion has similar, and good, reproducibility in the three studies that use it in some guise: CV*= 19.3 for S1 in Lee et al./Mille et al.; 17.8 for Adequacy Minor[10] in Popović/Popović & Belz; and 15.1 for Grammatical Correctness in Qader et al./Richter et al. Moreover, the study with the highest degree of reproducibility according to all measures (Lee et al./Mille et al.) obtained a comparatively large number of scores for each evaluated item, while also assessing a medium number of items per system. In contrast, the study with the lowest degree of reproducibility according to all measures (Mahamood et al./Mahamood) obtained a different number of scores for each evaluated item in the original and reproduction studies, while assessing a very small number (2) of items per system. We return to some of these aspects in the discussion section.

6 Discussion

There were considerable differences in evaluator cohorts between original and reproduction study in all four ReproGen study pairs. In Mahamood et al./Mahamood, the texts being evaluated were about progress towards getting a postgraduate degree (e.g.: *You haven't qualified for a postgraduate diploma. You have been awarded a postgraduate certificate instead. Average CAS results were achieved in CS5052, CS5038, CS5540, CS5548.*) Mahamood et al. (2007) asked postgraduate students to evaluate these texts, whereas Mahamood (2021) asked work colleagues to evaluate the texts. It is possible that students and non-students reacted differently to statements about degree progress, and that the students were much more familiar with terms such as 'postgraduate certificate' and 'CAS'.

There were also important differences in evaluator cohorts in Lee et al./Mille et al. and Qader et al./Richter et al.: in both cases, the reproduction cohort included people known to the authors personally who may have had more of an incentive to perform the task conscientiously and perhaps also to select higher scores. In the case of Lee et al./Mille et al., reproducibility was nevertheless good, whereas for Qader et al./Richter et al., it was less good.

Across all of our reproduction studies, there were differences in evaluators: age, recruitment, professional status, domain knowledge, background, etc. Such differences have the potential to impact reproducibility, but the picture from the four ReproGen studies was mixed, and further research is needed

[10] Assuming that grammatical errors account for much of minor translation adequacy issues, which is not certain.

to understand which characteristics were most important from this perspective. Knowing this would be very helpful in designing and interpreting experiments, as well as replicating them.

Both Track A reproduction studies contacted the original authors for additional information, highlighting the importance of original authors being willing to support reproduction studies of their work. It is clear from ReproGen'21 as well as other research (van der Lee et al., 2019; Howcroft et al., 2020; Belz et al., 2021) that we need a lot of information about evaluators and other aspects of evaluations in order to conduct reproduction studies, so it's essential that experimenters fill out a datasheet such as HEDS which conveys information in a standardised, comparable way.

A rarely mentioned aspect that should not be underestimated is that being willing to support a reproduction study of your work means being willing to take what some perceive as a substantial risk associated with having others publish assessments of the reproducibility of your work. Some authors are very uncomfortable with a reproduction study showing low reproducibility. In fact, one of the authors of a paper which had been the subject of a reproduction study that we wanted to include in our survey of reproduction studies (Belz et al., 2021) worried that the considerable gap in results would be interpreted as academic misconduct.[11]

Clearly there is a need for reproduction studies to be carried out in NLP. We need to know how reproducible different types of evaluation measures are, because measures with low reproducibility will result in unreliable results and unreliable conclusions based on them. Reproduction studies are the only way to know if/where we're going wrong in this sense. However, given prevailing sensitivities, it seems the right thing to do to conduct reproduction studies with the original authors' consent.

Reproduction studies are expensive and a lot work, and we were told by the five teams that registered for ReproGen but did not submit that these were the main reasons why they were ultimately not able to participate. Publication only provides so much of an incentive/motivation. Significant numbers of reproduction studies may only be feasible in the context of a funded project such as ReproHum,[13] where uniformity of approach can moreover be ensured and the number and type of reproduction studies conducted can be more directly controlled. We plan to run a second shared task next year, to further test the suitability of the shared task format for reproduction studies in NLP.

7 Conclusions

We first proposed the ReproGen shared task at Generation Challenges 2020[12] (Belz et al., 2020) and, taking into account feedback received, developed it into the shared task presented here, with the main track offering four original studies (sets of human evaluation results) for reproduction, and an open track inviting reproduction studies of own results.

Bearing in mind we had just one reproduction study for each original study available to us, and that as discussed we have to be cautious drawing conclusions based on sample sizes of 2, there are very few tentative first conclusions concerning reproducibility of human evaluation in NLG we have been willing to draw from ReproGen. We pointed out that the study with the highest degree of reproducibility obtained a comparatively large number of scores for each evaluated item, while also assessing a medium number of items per system. In contrast, the study with the lowest degree of reproducibility obtained a different number of scores for each evaluated item in the original and reproduction studies, while assessing a very small number (2) of items per system. We also observed that there was some evidence that the Grammaticality evaluation criterion has a comparatively good and stable degree of reproducibility.

When we read human evaluation results in NLG papers, unless there is an obvious red flag such as a very small number of evaluators, or evaluation items, we tend to trust those results more than metric results. Yet as we delve deeper into the reproducibility of our human evaluation results, it is beginning to become clear that, as a general assumption, this trust may be misplaced. More generally, that we need to do much more as a field to ensure that our human evaluation methods are fit for purpose, including in the sense that a rerun of an experiment will produce at least broadly similar results. With the ReproGen shared task, and the ReproHum project[13] which it is part of, we are aiming to make a contribution to this important goal.

[11] We therefore did not include the study in question in the published survey.

[12] INLG'20, Dublin.

[13] https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/V05645X/1

Acknowledgments

We thank the authors of the four original papers that were up for reproduction in Track A who bravely agreed to be our guinea pigs for this first shared task on reproducibility of evaluation measures in NLG. And of course the authors of the reproduction papers, the first batch of participants in a shared task on reproducibility of human evaluations 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.

Shubham Agarwal's PhD fees are supported by Adeptmind Inc., Toronto, Canada.

References

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

Anya Belz. 2021. Quantifying reproducibility in NLP and ML. *arXiv preprint arXiv:2109.01211*.

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 and Eric Kow. 2011. Discrete vs. continuous rating scales for language evaluation in NLP. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 230–235.

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.

Michael Cooper and Matthew Shardlow. 2020. CombiNMT: An exploration into neural text simplification models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5588–5594, Marseille, France. European Language Resources Association.

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.

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.

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.

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.

Saad Mahamood. 2021. Reproducing a comparison of hedged and non-hedged NLG texts. In *Proceedings of the 14th International Conference on Natural Language Generation*, Aberdeen, United Kingdom. Association for Computational Linguistics.

Saad Mahamood, Ehud Reiter, and Chris Mellish. 2007. A comparison of hedged and non-hedged nlg texts. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, pages 155–158.

Simon Mille, Thiago Castro Ferreira, Anya Belz, and Brian Davis. 2021. Another PASS - a reproduction study of the human evaluation of a football report generation system. In *Proceedings of the 14th International Conference on Natural Language Generation*, Aberdeen, United Kingdom. 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.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2020. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *CoRR abs/2003.12206*.

Maja Popović. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maja Popović and Anya Belz. 2021. A reproduction study of an annotation-based human evaluation of MT outputs. In *Proceedings of the 14th International Conference on Natural Language Generation*, Aberdeen, United Kingdom. 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.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Christian Richter, Yanran Chen, and Steffen Eger. 2021. TUDA-reproducibility @ reprogen: Replicability of human evaluation of text-to-text and concept-to-text generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, Aberdeen, United Kingdom. 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. 2021. The human evaluation datasheet 1.0: A template for recording details of human evaluation experiments in NLP. *CoRR*, abs/2103.09710.

Koustuv Sinha, Joelle Pineau, Jessica Forde, Rosemary Nan Ke, and Hugo Larochelle. 2020. NeurIPS 2019 Reproducibility Challenge. *ReScience C*, 6(2):#11.

Measurand	Orig study	Repro study	% change	CV*
Original study = van der Lee et al. (2017); reproduction study = Mille et al. (2021):				
Stance identification Accuracy, PASS system	91	96.75	6.32	6.107
Mean Clarity, 0..6†, PASS system	4.64	5.3	10.7	13.193
Mean Fluency, 0..6†, PASS system	4.36	5.14	13.55	16.372
Original study = Mahamood et al. (2007); reproduction study = Mahamood (2021):				
Strength of preference for style A vs. B (0..6†)				
Native speakers, Scenario 1	1.58	0.8	-49.37	65.35
Native speakers, Scenario 2	0.93	1.6	72.04	52.806
Fluent speakers, Scenario 1	3.09	1.0	-67.64	101.894
Fluent speakers, Scenario 2	3.45	1.67	-51.59	69.323
Original study = Popović (2020); reproduction study = Popović and Belz (2021):				
Comprehension Minor, % words with errors				
Bing	16.0	16.8	+5	4.86
Google	11.2	15.0	+33.93	28.92
Amazon	12.0	16.7	+39.17	32.65
Comprehension Major, % words with errors				
Amazon	7.6	10.2	+34.21	29.13
Bing	15.1	22.3	+47.68	38.38
Google	7.1	11.5	+61.97	47.17
Adquacy Minor, % words with errors				
Google	10.5	11.7	+11.43	10.78
Amazon	11.4	13.1	+14.91	13.84
Bing	17.0	12.7	-25.29	28.87
Adequacy Major, % words with errors				
Google	7.0	9.7	+38.57	32.24
Amazon	6.5	9.5	+46.15	37.39
Bing	13.2	21.2	+60.61	46.37
Original study = Qader et al. (2018); reproduction study = Richter et al. (2021):				
Mean Information Coverage, 0..4†				
Reference	2.1	2.9	38.1	31.904
C2T	1.9	1.5	-21.05	23.459
C2T_char	1.3	2.0	53.85	42.297
C2T+pg	1.3	1.6	23.08	20.628
C2T+pg+cv	1.7	2.0	17.65	16.168
T2T+pg	0.8	1.6	100	66.467
T2T+pg+cv	1.3	1.9	46.15	37.388
Mean Non-redundancy of Information, 0..4†				
Reference	3.6	3.1	-13.89	14.881
C2T	1.9	2.8	47.37	38.183
C2T_char	2.9	1.8	-37.93	46.668
C2T+pg	3.5	3.2	-8.57	8.928
C2T+pg+cv	2.9	3.1	6.9	6.647
T2T+pg	2.3	2.5	8.7	8.308
T2T+pg+cv	2.8	3.1	10.71	10.139
Mean Semantic Adequacy, 0..4†				
Reference	2.9	2.9	0	0
C2T	2.3	1.6	-30.43	35.79
C2T_char	1.8	2.1	16.67	15.339
C2T+pg	3.0	1.9	-36.67	44.764
C2T+pg+cv	2.6	2.9	11.54	10.876
T2T+pg	1.9	1.7	-10.53	11.078
T2T+pg+cv	1.4	1.8	28.57	24.925
Mean Grammatical Correctness, 0..4†				
Reference	3.2	3.0	-6.25	6.432
C2T	2.6	2.2	-15.38	16.617
C2T_char	2.0	2.5	25	22.156
C2T+pg	3.3	2.8	-15.15	16.344
C2T+pg+cv	3.2	3.1	-3.13	3.165
T2T+pg	2.7	3.0	11.11	10.495
T2T+pg+cv	2.5	3.4	36	30.417

Table 4: Overview of results from ReproGen'21 reproduction studies: measurand, measured value in original study, measured value in reproduction study, percentage change (in/decrease), and coefficient of variation (CV*). † the original scale was shifted to start from 0.