Non-Repeatable Experiments and Non-Reproducible Results: The Reproducibility Crisis in Human Evaluation in NLP

Anya Belz[a,b] **Craig Thomson**[b] **Ehud Reiter**[b] **Simon Mille**[a]

[a]ADAPT, Dublin City University
Dublin, Ireland
{anya.belz,simon.mille}@adaptcentre.ie

[b]University of Aberdeen
Aberdeen, UK
{c.thomson,e.reiter}@abdn.ac.uk

Abstract

Human evaluation is widely regarded as the litmus test of quality in NLP. A basic requirement of all evaluations, but in particular where used for meta-evaluation, is that they should support the same conclusions if repeated. However, the reproducibility of human evaluations is virtually never queried in NLP, let alone formally tested, and their repeatability and reproducibility of results is currently an open question. This paper reports our review of human evaluation experiments published in NLP papers over the past five years which we assessed in terms of (i) their ability to be rerun, and (ii) their results being reproduced where they can be rerun. Overall, we estimate that just 5% of human evaluations are repeatable in the sense that (i) there are no prohibitive barriers to repetition, and (ii) sufficient information about experimental design is publicly available for rerunning them. Our estimate goes up to about 20% when author help is sought. We complement this investigation with a survey of results concerning the reproducibility of human evaluations where those are repeatable in the first place. Here we find worryingly low degrees of reproducibility, both in terms of similarity of scores and of the findings supported by them. We summarise what insights can be gleaned so far regarding how to make human evaluations in NLP more repeatable and more reproducible.

1 Introduction

Human evaluation is widely seen as the most reliable form of evaluation in NLP. The traditional view in the field, here expressed for MT, is that "automatic measures are an imperfect substitute for human assessment of translation quality" (Callison-Burch et al., 2008). Numerous papers have reported meta-evaluations of metrics in terms of correlation with human judgments (Belz and Reiter, 2006; Espinosa et al., 2010; Hashimoto et al., 2019; Clark et al., 2019; Sellam et al., 2020). However, recently several papers have highlighted issues arising from lack of standardisation (Belz et al., 2020), incomplete details reported for evaluation design (Howcroft et al., 2020), and poor experimental standards (van der Lee et al., 2019). In this paper, we address issues that intersect with all of these.

Our starting premise is that in order to act as litmus test of quality, human evaluations need to be able to be relied upon to produce the same results, at least in the sense of supporting the same conclusions, when run multiple times. This ought to be a low-threshold requirement, but is in fact very rarely assessed at all, let alone routinely established for new evaluation methods. Inter-evaluator agreement is more commonly assessed, but falls far short of establishing whether an experiment when repeated produces similar results and/or supports similar findings. Our aim in the work reported here[1] is establishing the reproducibility, or otherwise, of current human evaluation practices, in order to provide evidence-based indications regarding how they can be improved, thereby going beyond recent opinion-based recommendations regarding better practice.

This paper makes five main contributions: (i) an annotation scheme capturing experimental properties playing a role in repeatability and reproducibility (Section 2 and Table 1); (ii) an assessment of the *repeatability* of human evaluation experiments in NLP (Section 2); (iii) a state-of-the-field assessment of the *reproducibility* of human evaluations in NLP (Section 3); (iv) the dataset of paper details and annotations our analyses are based on;[2] and (v) evidence-based recommendations regarding how to improve the repeatability and reproducibility of human evaluations in NLP (Section 4).

We use the terms repeatability and reproducibility as follows. ***Repeatable*** is a property of experiments meaning *able to be repeated with identical experimental design*. ***Reproducible*** is a property of

[1]Part of the ReproHum project: https://reprohum.github.io/

[2]To be released via the project website: https://reprohum.github.io/

Findings of the Association for Computational Linguistics: ACL 2023, pages 3676–3687
July 9-14, 2023

property	definition
dataset	Which dataset was used?
quality criterion	Which quality criterion is being assessed?
language(s)	Which language(s) were evaluated?
task	Which NLP task best describes the experiment?
participant type	What type of participants were recruited? (crowd-source, student, etc)
intrinsic or extrinsic	Is the evaluation intrinsic or extrinsic?
absolute or relative	Is the evaluation absolute or relative?
objective or subjective	Is the evaluation objective or subjective?
total participants	How many total participants are there?
total items	How many items in total?
participants per item	How many participants per item?
items per participant	How many items per participant?
training session	Do participants take part in a training session?
participant instructions	Are participants given instructions or shown worked examples? (beyond basic instructions on using the interface or basic criterion definitions)
quality criterion definitions	Are participants shown definitions for any criterion they are asked to evaluate?
participant practice task	Do participants complete practice tasks prior to the main experimental tasks?
participant custom qualification	Are participants required to pass a custom qualification exercise?
participant expertise (controlled?)	Does the experiment control for participant expertise?
participant expertise (self report?)	Is any reported expertise self-reported (with no interaction with researchers, i.e. on MTurk or a double-blind study)?
participant expertise (type?)	What type of expertise do participants have?
native speakers (control?)	Does the experiment include control for recruiting only native speakers?
native speakers (self report?)	Is any reported native speaking self-reported (with no interaction with researchers, i.e. on MTurk or a double-blind study)?

Table 1: Properties in experiment annotation scheme.

evaluations, meaning *producing the same results and/or findings when run multiple times*.

2 Repeatability of Human Evaluation Experiments

In this section, we describe (Section 2.1) our 4-stage process for assessing human evaluations in terms of *repeatability* as a precondition for inclusion in a coordinated set of reproductions. As part of this process, we annotated papers and then experiments with evaluation properties, and we examine what these reveal. Because the final stage of this selection process introduced non-systematic selection (to meet the needs of the coordinated studies design), we also verify our findings on a separate, randomly selected subset of papers (Section 2.2). For an overview of the selection/filtering process, see the flow diagram in Appendix A.

2.1 Identifying repeatable experiments

Selection procedure. To start, we extracted all papers containing the key phrases "human evaluation" and "participants" from TACL and the ACL main conference in the ACL Anthology (177). We included papers from 2018 to 2022 inclusive.[3] We manually checked and excluded papers that did not report a new human evaluation of system outputs.

Paper-level properties. In the second stage we annotated seven paper-level properties including language(s), number of systems, dataset and participants (for details, see Belz et al., 2023). During the annotation process, we excluded papers that had prohibitive barriers to reproduction, which meant those that we estimated to have cost >USD 2,000)[4], and/or that had a longitudinal design, and/or that used highly specialised experts as evaluators such as doctors[5]. This left 116 papers, of which 29 are from TACL and 87 from ACL.

Experiment-level properties. We then split each paper into the experiments it reports and started annotating each experiment with our fine-grained annotation scheme (Table 1; for additional details see Appendix B). At this point we estimated we had enough information to complete the annotations in the case of just 5% of our papers. We therefore started contacting authors to obtain the missing information. Following the prolonged contacting process (for details see Belz et al., 2023, and Appendix C), we obtained the requested information for just 20 papers (containing 28 experiments). Using both the publicly available and author-provided

[3]The search was performed in mid-2022 and does include ACL 2022 but not all TACL papers from 2022.

[4]Using a rough estimate based on numbers of items evaluated, evaluators, and evaluation platform. 6 papers were excluded for this reason.

[5]Only 9 papers were excluded for these reasons. Most excluded papers simply did not report a human evaluation of system outputs.

information, we were able to collate property values to the extent shown in Table 3: 20 experiments had no unclear properties, and 8 had one or more.

That we were able to find clear properties for 20 of the 28 experiments in Table 3 does not indicate that these experiments could definitely be recreated, just that we have the minimal level of information required to attempt recreation. That we can only clear this first hurdle for 17% of the 115[6] papers we started with is alarming.

Bugs, errors and flaws. Moreover, in the process of collating and checking experiment details, we found several types of issues that in some cases called into question whether they should be repeated at all, for ethical and/or scientific reasons (for details see Belz et al., 2023).

2.2 Verification on random subset of papers

In order to verify the above finding that only 5% of papers are repeatable from publicly available information, we sampled a new batch of papers from an expanded set of 631 ACL, TACL and EMNLP papers that matched the keyword search, and did not fail any of our inclusion tests as above.

We annotated the 26 experiments reported in these 20 randomly sampled papers using the same procedure as in Section 2.1, except that we only used information that was publicly available either from the paper, supplementary material, or hyperlinks in the paper, e.g., a GitHub repository. In particular we tried to find the system outputs that were shown to participants, and the interface, form, or document that participants completed.

We found the above information for just 5% of papers, confirming our estimate from Section 2.1. Three papers made either just the interface or just the system outputs available. Table 2 shows the number of experiments out of all 26 where a given property was clear, for all properties in our annotation scheme. It is clear from the numbers in the table that very basic information such as number and type of participants is very often not findable.

3 Reproducibility of Results from Human Evaluations

To complement the assessment of the repeatability of human evaluations in NLP above, here we look at the *reproducibility* of results, as collated from recent reproduction studies. We examine similarity in system-level scores between original and reproduction studies (Section 3.1), and assess whether scores support the same conclusions which can be the case even for dissimilar scores (Section 3.2).

3.1 Similarity of scores

Table 5 provides an overview of reproducibility results from reproduction studies of human system quality evaluations performed as part of the REPROLANG (Branco et al., 2020), ReproGen 2021 (Belz et al., 2021a), and ReproGen 2022 (Belz et al., 2021b) shared tasks. We exclude evaluations based on text annotation where a single overall aggregated score per system was not computed.

Column 1 identifies the original and reproduction study and the evaluation criteria assessed. The last two columns show the corresponding mean study-level and mean criterion-level coefficients of variation (CV*) (Belz et al., 2022), and rank preservation, respectively. The columns in between show seven properties of each study/criterion, as per the HEDS datasheet (Shimorina and Belz, 2022); column headings identify HEDS question number (see table caption for explanation).

3.2 Confirmation of conclusions

Another perspective on reproducibility is whether the same conclusions can be drawn from two evaluations. Table 4 assesses the (dis)similarity of ranks between the pairs of original and reproduction ex-

property	#clear
dataset	26
quality criterion	26
language(s)	26
task	26
participant type	14
intrinsic or extrinsic	26
absolute or relative	26
objective or subjective	26
total participants	14
total items	22
participants per item	19
items per participant	13
training session	1
participant instructions	3
quality criterion definitions	5
participant practice task	2
participant custom qualification	1
participant expertise (controlled?)	13
participant expertise (self reported?)	13
participant expertise (type?)	10
native speakers (control?)	10
native speakers (self reported?)	10

Table 2: Number of experiments out of 26 for which a given property was clear (**random sample of 20 papers using publicly available information only**).

[6] 116 minus one paper we excluded after receiving a response from the author.

property	#clear
dataset	28
quality criterion	28
language(s)	28
task	28
participant type	26
intrinsic or extrinsic	28
absolute or relative	28
objective or subjective	28
total participants	23
total items	26
participants per item	28
items per participant	24
training session	24
participant instructions	23
quality criterion definitions	25
participant practice task	24
participant custom qualification	25
participant expertise (controlled?)	28
participant expertise (self-report?)	28
participant expertise (type?)	28
native speakers (controlled?)	24
native speakers (self-report?)	24

Table 3: Number of experiments out of 28 for which a given property was clear (**non-random set of 20 papers where authors provided missing information**).

Original / reprod.	Evaluation criterion	ρ	r
Nisioi et al. / Popovic et al.	Simplicity	0.73	0.77
Qader et al. / Richter et al.	Info Coverage	0.29	0.57
	Info Non-redundancy	0.499	0.33
	Semantic Adequacy	0.396	0.52
	Gram. Correctness	0.196	0.32
Nisioi et al. / Cooper & Shardlow	Grammaticality	-1	-1
	Meaning Preserv.	-1	-1
Popovic / Popovic & Belz	Compreh. Minor	1	0.67
	Compreh. Major	0.5	0.99
	Adequacy Minor	0.5	0.36
	Adequacy Major	1	0.999
Mahamood et al. / Mahamood	Preference (native)	-1	-1
	Preference (fluent)	1	1

Table 4: Spearman's ρ as an indication of how closely matched system ranks are between original and reproduction studies (Pearson's r for reference).

periment from Table 5. A clear picture emerges: scores in reproductions correlate positively (Pearson's r) with those in original studies, but correlations are not strong in most cases. Most importantly, system ranks are not the same as in the original experiment in any of the reproductions, although for individual evaluation criteria they are the same (Spearman's ρ=1) in three cases.

4 Discussion

When corresponding with authors to find missing information (Section 2.1), and when trying to find information from publicly available sources (Section 2.2), properties were often not obtainable for similar reasons. High-level properties such as the dataset, task, and language, could usually be found in the paper. The total number of items was usually available, but the relationship between participants and items was not. Information regarding recruitment of participants, as well as what they saw and did during the experiment almost always required additional information from the author. If authors were to make the files they used for the experiment, and a record of how these were processed (including the way they were presented to participants), then it would go a long way towards making the recreation of more experiments possible.

Repeatability, in the sense of being able to be repeated, is a basic requirement of all scientific experiments, perhaps most importantly as a prerequisite to independent verification through reproduction: "An experimental result is not fully established unless it can be independently reproduced" (ACM, 2020). It is therefore of concern in and of itself that the large majority (95%) of human evaluations in NLP is not repeatable from publicly available information (Section 2.2). This is further compounded by our finding (Section 2.1) that even with considerable effort (up to three emails to first and if necessary other authors) to obtain missing information to enable repetition, 80% of experiments remain non-repeatable.

Finally, where we were able to obtain and review all information needed for a repetition, we found multiple reporting mistakes, errors in scripts, and ad-hoc manual interference in live experiments that call into question for scientific and/or ethical reasons whether experiments should be repeated.

Our analysis of reproduction results (Table 5) showed that for the simplest binary output categorisation task, a good degree of reproducibility could be achieved (CV* = 6.11), but for most of the other, more cognitively complex, evaluations, degree of reproducibility was poor. Most significantly, the same set of conclusions could not be drawn regarding ranks of systems evaluated in any of the reproductions at the experiment level.

We would argue that we urgently need to (i) improve the repeatability of human evaluation experiments by making available publicly, as standard, full information about how the experiment was conducted, in sufficient detail to enable others to re-run it; (ii) test the results reproducibility

Original / reproduction study, measurand	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* ↓	ranks same?
Nisioi et al. / Popovic et al., Simplicity	70	3/3	-2,-1,0,1,2	DQE	Feature	Both	RtI	2	*8.98*	no
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	n/a
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	
Qader et al. / Richter et al.									*22.16*	–
Information Coverage	30	19/19	1–5	DQE	Corr	Cont	RtI	1	34.04	no
Information Non-redundancy	30	19/19	1–5	DQE	Good	Cont	iiOR	1	19.11	no
Semantic Adequacy	30	19/19	1–5	DQE	Corr	Cont	iiOR	1	20.4	no
Grammatical Correctness	30	19/19	1–5	DQE	Corr	Form	iiOR	1	15.09	no
Nisioi et al. / Cooper & Shardlow									*25.55*	–
Grammaticality	70	3/5	1–5 / 1–10	DQE	Corr	Form	iiOR	?	25.01	no
Meaning Preservation	70	3/5	1–5 / 1–10	DQE	Corr	Cont	RtI	?	26.08	no
Popović / Popović & Belz									*29.22*	–
Comprehension Minor	557,	7/7	2 labels	Anno	Good	Both	iiOR	2	22.14	yes
Comprehension Major	279,	7/7		Anno	Good	Both	iiOR	2	38.23	no
Adequacy Minor	467	7/7	3 labels	Anno	Corr	Cont	RtI	2	17.83	no
Adequacy Major		7/7		Anno	Corr	Cont	RtI	2	38.67	yes
Mahamood et al. / Mahamood, Binary Preference Strength	2†	25‡/11	-3..+3	RQE	Good	Both	EFoR	25/11	*72.34*	no

Table 5: Overview of reproducibility results from existing reproduction studies in terms of (mean) CV* and rank preservation (last two columns). Evaluations are characterised in terms of some properties from HEDS datasheets: 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.

of new evaluation methods prior to running full evaluation experiments with them; (iii) standardise evaluation methods, especially measurand (evaluation criterion) and measurement procedure, so that the reproducibility of each, once established, does not have to be tested every time. The worrying levels of errors and flaws in reporting and design we found can be in part addressed through standardisation and establishing reproducibility for standardised methods, but will also require a shift in expectations and awareness of how to conduct good quality human evaluations for NLP.

5 Conclusion

NLP needs human evaluation as a litmus test of quality, including as a reliable reference for meta-evaluating other types of evaluation. In order to play this role, human evaluation needs to be verifiably reliable, and that includes being reproducible; in order to assess the reproducibility of results, we need to be able to repeat an experiment. However, our results showed that current human evaluations have very poor repeatability (we estimated that just 5% do not have prohibitive barriers to being repeated, and can be re-run without recourse to non-public information), and where we are able to repeat human evaluations, the growing number of results from human evaluation reproduction studies show that they have low degrees of reproducibility of both scores and conclusions. We derived recommendations for making human evaluations in NLP more repeatable and more reproducible, something that we surely need to do if we are to continue treating them as our most trusted assessment of system quality.

Acknowledgements

The ReproHum project is funded by EPSRC grant EP/V05645X/1. We would like to thank all authors who took the time to respond to our requests for information. We would also like to thank all our collaborators at the ReproHum partner labs: Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Jackie Cheung, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondrej 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, Ondrej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, Diyi Yang.

Limitations

Those of our findings that are based on information obtained from authors are necessarily limited in that they do not reflect information that might have been obtained from authors who did not respond.

Moreover, we selected our initial set of papers via search with key phrase "human evaluation." While this phrase is very commonly used to refer to non-automatic forms of evaluation, there is a chance that we may have missed papers because they used a different term.

Conclusions based on our randomly selected sample of 26 experiments, and non-random sample of 28 experiments, are limited by their sample size in terms of how representative they are likely to be of current human evaluations in NLP more generally.

Ethics Statement

As a paper that meta-reviews other academic publications, the present paper can be considered low-risk. Over and above collating information from publications, we annotated papers, analysed results and obtained descriptive statistics from annotations. In Section 2.1, we summarise the flaws, bugs and errors we found in experiments we were preparing for reproduction studies, We decided not to cite the papers where we found these, because the important information was that such issues occur, not which researchers were responsible for them. See also the responsible NLP research checklist completed for this paper.

References

ACM. 2020. Artifact review and badging Version 1.1, August 24, 2020. https://www.acm.org/publications/policies/artifact-review-and-badging-current.

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, Maja Popovic, 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, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

Anya Belz, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021a. The ReproGen shared task on reproducibility of human evaluations in NLG: Overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 249–258, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Anya Belz, Anastasia Shimorina, Maja Popovic, and Ehud Reiter. 2021b. The ReproGen shared task on reproducibility of human evaluations in NLG: Overview and results. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 249–258.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Jackie Cheung, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondrej 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, Ondrej 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.

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 Twelfth Language Resources and Evaluation Conference*, pages 5539–5545, Marseille, France. European Language Resources Association.

Chris Callison-Burch, Cameron Shaw Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the third workshop on statistical machine translation*, pages 70–106.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

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

Dominic Espinosa, Rajakrishnan Rajkumar, Michael White, and Shoshana Berleant. 2010. Further meta-evaluation of broad-coverage surface realization. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 564–574.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. 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.

Saad Mahamood. 2021. Reproducing a comparison of hedged and non-hedged NLG texts. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 282–285, Aberdeen, Scotland, UK. 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, Saarbrücken, Germany. DFKI GmbH.

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*, pages 286–292, Aberdeen, Scotland, UK. 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. 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*, pages 293–300, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Maja Popović, Sheila Castilho, Rudali Huidrom, and Anya Belz. 2022. A reproduction study of an annotation-based human evaluation of MT outputs. In *Proceedings of the 15th International Conference on Natural Language Generation*, Waterville, Maine, USA and virtual meeting. 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.

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*, pages 301–307, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

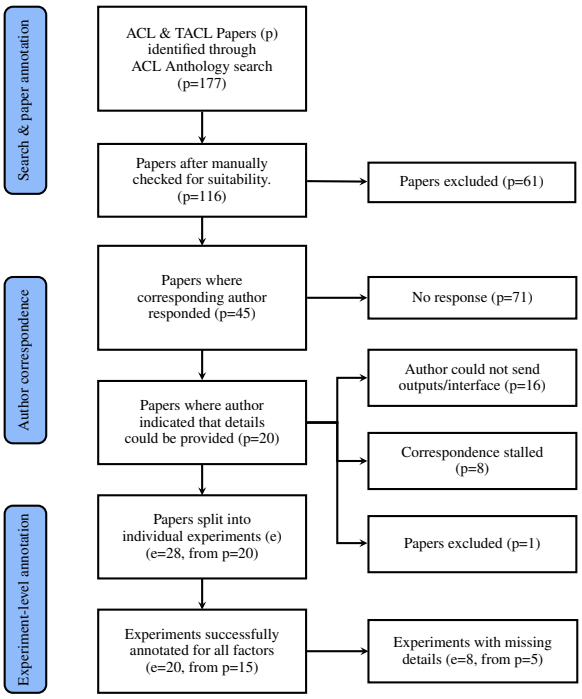
Anastasia Shimorina and Anya Belz. 2022. The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.

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.

A Flow Diagram of Paper Selection/Filtering Process

The following diagram shows the steps in paper selection/filtering process (reproduced from Belz et al. (2023), for ease of reference):



B Details of evaluation experiment properties

All of the property names and values from our detailed annotations are listed below, along with descriptions of what was recorded for each property:

1. Specific data sets used;
2. Specific evaluation criteria names used; the criterion names as stated in the paper if possible, otherwise a criterion name that represents what is being assessed.
3. System languages; the language(s) used by the system as either input or output.
4. System task; the NLP task that the system is tackling. Values from the 28 experiments were cross-lingual summarisation, data-to-text generation, definition generation with controllable complexity, dialogue summarisation, dialogue turn generation, explanation generation, fact-check justification generation, machine translation error prediction, prompted generation, question generation, question-answer generation, referring expression generation, simplification, summarisation, text to speech.
5. Evaluator type; the type of evaluator, values included colleagues, commercial in-house evaluators, crowd-sourced, mix of author and colleague, mix of colleague and students, professional, student.
6. Evaluation modes (**?**):
   (a) Intrinsic vs. extrinsic;
   (b) Absolute vs. relative;
   (c) Objective vs. subjective.
7. Number of participants; the total number of unique participants that took part in the study,
8. Number of items evaluated; in the case of an absolute evaluation this is one system output. In the case of a relative evaluation, it refers to the set of outputs, e.g., a pair, that is being compared.
9. How many participants evaluated each item; for some experiments, this varied.
10. How many items were evaluated by each participant; for some experiments, this varied. In particular, for the 13 of 28 experiments that were crowd-sourced, 5 were known integers,

4 varied, and 4 could not be determined (we suspect these also varied).

11. Were training and/or practice sessions provided for participants; see the discussion below.
12. Were participants given instructions? Were they given definitions of evaluation criteria; see the discussion below.
13. Were participants required to have a specific expertise? If so, what type, and was this self-reported or externally assessed?; see the discussion below.
14. Were participants required to be native speakers? If so, was this self-reported or externally assessed?; For the first part we used the options yes, no, crowd-source region filters, and in one case that the experiment was performed with students at a university where the language was native. The latter two are inherently self-reported, although with some limited control by the researchers. Only for one of the experiments with native speakers did the researchers indicate that they had confirmed this, all others were self-reports.
15. How complex was the evaluation task (low, medium, high); assessment by authors of this paper.
16. How complex was the interface (low, medium, high); assessment by authors of this paper.

Classifying the type of participant, training, instruction, and expertise was very difficult. Firstly, not all experiments necessarily require detailed instructions but setting a threshold beyond which instructions become non-perfunctory is difficult. The same is true for training. In the end, we decided to record whether there non-perfunctory training, instruction, practice, or criterion definition.

Expertise was also difficult to classify. Some papers would have originally reported 'expert annotators', but following our queries stated participants were graduate students or colleagues. Such participants were often called 'NLP experts'. In the end, we considered participants to be expert if the authors of the original study indicated that they were.

C Process for contacting authors

When we contacted authors of papers we followed a standard procedure. We considered the corresponding author to be the first author of the paper, unless a different corresponding author was explicitly stated. First they were sent the following email:

> Dear **«NAME»**,
>
> The ReproHum project at the University of Aberdeen is running a multi-lab study where over 20 partner labs from across the world will be reproducing human evaluation experiments from NLP papers. The project is being led by Prof. Anya Belz, with Prof. Ehud Reiter as co-investigator, and myself as a research assistant.
>
> To create a shortlist of papers to reproduce, we looked for papers containing human evaluations, at high-profile conferences such as **«VENUE»**. We identified your paper "**«TITLE»**" from **«VENUE» «YEAR»** as a candidate for inclusion in our study. If included, the human evaluation that was performed for the paper would initially be reproduced by 2 different labs. One of our main objectives is to identify types of human evaluation that are associated with higher degrees of reproducibility so that the NLP community can then use this information to select the most appropriate methods for their studies.
>
> We are writing to you today to ask if you can provide us with more information about your experiment to enable us to reproduce it under conditions that are as close to the original as possible. We are particularly hoping that you can provide the system outputs and questions that were shown to participants.
>
> We would be most grateful if you could initially confirm that you are able to send us (links to) the below information (for each human evaluation that is reported in the paper):
>
> 1. The system outputs that were shown to participants.
> 2. The interface, form, or document that participants completed; the exact document or form that was used would be ideal.
> 3. Details on the number and type of participants (students, researchers, Mechanical Turk, etc.) that took part in the study.
> 4. The total cost of the original study.
>
> If you are able to provide the above information, we would be grateful if you could also confirm how soon this would be possible.
>
> If you have any questions, please contact us.
>
> With best regards,
>
> Anya, Ehud, and Craig
>
> Project web page: https://reprohum.github.io

If there was no response the above email, they were sent a second email with only minor adjustments to reflect that we had tried to contact them previously. A third email was sent in cases where we still had no response. At least one week passed between each email sent to an author. The first two emails were sent from the academic email account of a research assistant, although addressed from the

whole project team. The third email was sent by a professor, and whilst this did elicit a small number of responses, most came from the first two emails. In the event that email addresses were no longer valid, we searched for a more recent email for the author, primarily by checking their most recent papers. In the event that we could not find any email address for an author, we attempted to contact the next author in the same way. We were able to find a working email address for one author from all bar one paper. Most were sent using a mail merge, although some were aggregated and sent manually, in cases where one author had many papers.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
First non-numbered section after conclusions.

☑ A2. Did you discuss any potential risks of your work?
Risk is largely not applicable, but we mention deciding against naming authors of papers with flaws/bugs/errors to avoid reputational damage.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
See abstract, introduction (Section 1) and conclusions.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
All papers are cited throughout.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We mention finding non-anonymised datasets in public resources and taking that as a reason not to use them.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.