Missing Information, Unresponsive Authors, Experimental Flaws: The Impossibility of Assessing the Reproducibility of Previous Human Evaluations in NLP

Anya Belz[a,b] (anya.belz@adaptcentre.ie), **Craig Thomson**[b], **Ehud Reiter**[b], **Gavin Abercrombie**[8], **Jose M. Alonso-Moral**[17], **Mohammad Arvan**[16], **Anouck Braggaar**[13], **Mark Cieliebak**[20], **Elizabeth Clark**[6], **Kees van Deemter**[19], **Tanvi Dinkar**[8], **Ondřej Dušek**[9], **Steffen Eger**[1], **Qixiang Fang**[19], **Mingqi Gao**[11], **Albert Gatt**[19], **Dimitra Gkatzia**[4], **Javier González-Corbelle**[17], **Dirk Hovy**[2], **Manuela Hürlimann**[20], **Takumi Ito**[10], **John D. Kelleher**[12], **Filip Klubička**[12], **Emiel Krahmer**[13], **Huiyuan Lai**[7], **Chris van der Lee**[13], **Yiru Li**[7], **Saad Mahamood**[14], **Margot Mieskes**[15], **Emiel van Miltenburg**[13], **Pablo Mosteiro**[19], **Malvina Nissim**[7], **Natalie Parde**[16], **Ondřej Plátek**[9], **Verena Rieser**[8], **Jie Ruan**[11], **Joel Tetreault**[3], **Antonio Toral**[7], **Xiaojun Wan**[11], **Leo Wanner**[18], **Lewis Watson**[4], **Diyi Yang**[5]

[a]ADAPT/DCU, Ireland; [b]University of Aberdeen, UK; [1]Bielefeld University, Germany; [2]Bocconi University, Italy; [3]Dataminr, US; [4]Edinburgh Napier University, UK; [5]Georgia Tech, US; [6]Google Research, US; [7]Groningen University, Netherlands; [8]Heriot-Watt University, UK; [9]Charles University Prague, Czechia; [10]Tohoku University, Japan; [11]Peking University, China; [12]Technological University Dublin, Ireland; [13]Tilburg University, Netherlands; [14]trivago, Germany; [15]University of Applied Sciences Darmstadt, Germany; [16]University of Illinois Chicago, US; [17]Universidade de Santiago de Compostela, Spain; [18]Universitat Pompeu Fabra, Spain; [19]Utrecht University, Netherlands; [20]Zurich University of Applied Sciences, Switzerland

Abstract

We report our efforts in identifying a set of previous human evaluations in NLP that would be suitable for a coordinated study examining what makes human evaluations in NLP more/less reproducible. We present our results and findings, which include that just 13% of papers had (i) sufficiently low barriers to reproduction, and (ii) enough obtainable information, to be considered for reproduction, and that all but one of the experiments we selected for reproduction was discovered to have flaws that made the meaningfulness of conducting a reproduction questionable. As a result, we had to change our coordinated study design from a reproduce approach to a standardise-then-reproduce-twice approach. Our overall (negative) finding that the great majority of human evaluations in NLP is not repeatable and/or not reproducible and/or too flawed to justify reproduction, paints a dire picture, but presents an opportunity for a rethink about how to design and report human evaluations in NLP.

1 Introduction

There is increasing awareness in Natural Language Processing (NLP) that reproducibility of results, most particularly of results from system evaluations, matters greatly, and that currently the field does not assess reproducibility of results rigorously enough, and lacks a common approach to it. Recent work has made progress particularly with respect to automatic evaluation (Pineau, 2020; Whitaker, 2017), but reproducibility of human evaluation, widely considered the litmus test of quality in NLP, has received less attention. It could be argued that if it is not known how reproducible human evaluations are, it is not known how reliable they are; and if it is not known how reliable they are, then it is not known how reliable automatic evaluations meta-evaluated against them are either.

The work reported in this paper forms part of the ReproHum project[1] in which our aim is to build on existing work on recording properties of human evaluations datasheet-style (Shimorina and Belz, 2022), and assessing how close results from a reproduction study are to the original study (Belz et al., 2022), to investigate systematically what factors make a human evaluation more—or less—reproducible. In this paper, we present the findings from our work on the project so far which necessitated a rethink of our entire approach to designing such an investigation.

Section 2 outlines our motivation for carrying

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

The Fourth Workshop on Insights from Negative Results in NLP, pages 1–10

out a multi-lab multi-test (MLMT) study of factors affecting reproduciblity in NLP, and our original design for the study. Section 3 describes our paper selection, annotation and filtering process which yielded a surprisingly small number of candidate papers for reproduction. In Section 4 we describe the numerous further issues with original evaluation studies we encountered in the process of setting up reproductions of them with partner labs. Section 6 summarises our negative findings regarding the infeasibilty of assessing the reproducibility of previously conducted human evaluations in NLP as they are, and outlines the changes to our multi-lab multi-test study necessitated by the findings.

2 Motivation and Overall Study Design

Individual studies can tell us how close a reproduction study's results are to those in the original study. A large number of such studies can show general tendencies regarding what kinds of evaluations have better reproducibility. However, we do not currently have a large number of reproduction studies in NLP and because of their cost and lack of appeal, this is unlikely to change. Moreover, accumulations of individual studies do not provide the conditions in which the effect size and significance of specific factors on reproducibility, and interactions between them, can be measured.

To create such conditions, a controlled study of equal numbers of reproductions with and without factors of interest is needed. Moreover, we know from existing work (Belz et al., 2022; Huidrom et al., 2022) that different reproductions of the *same* original work can produce very different results. Finally, while it is instructive to test for reproducibility under identical conditions, it is also of interest to test how far good reproducibility can stretch – e.g. is reproducibility affected by replacing, say, a 7-point quality scale with a 5-point one.

A study of factors that increase/decrease reproducibility therefore needs to (i) conduct more than one reproduction of each original study, (ii) carried out by a good mix of different teams, and to (iii) incorporate multiple rounds with decreasing similarity of conditions. The steps in setting up such a study would be as follows:

1. Identifying candidate evaluation experiments from which to select experiments with balanced factors to include in the MLMT study;
2. Recording properties of evaluation experiments to make it possible to select factors and control for them;
3. Selecting factors to control for and corresponding subsets of experiments; and
4. Carrying out reproductions for the selected evaluation studies and factors.

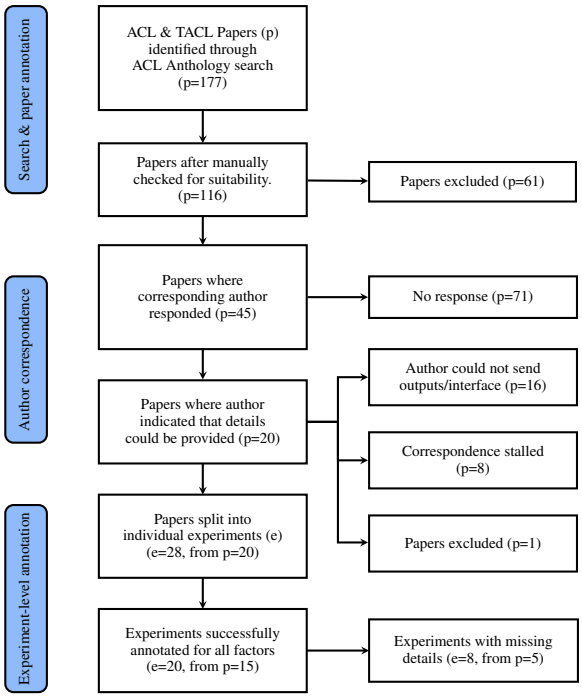

Figure 1: Flow diagram of the paper selection process, showing the decreasing number of papers that were suitable as more information was sought.

We describe Steps 1 and 2 in Sections 3.1 and 3.2, Step 3 in 3.3, and Step 4 up to the point where we aborted the original study design in Section 4.

3 Selection and Assessment of Candidate Evaluation Experiments

Figure 1 shows the selection and annotation process in the form of a flow diagram showing the decreasing number of remaining papers/experiments. The first step was to conduct a search on the ACL Anthology for papers published in ACL (main conference) or TACL in the 2018–2022 period[2] which included the phrases "human evaluation" and "participants;" we found 177 such papers.

3.1 High-level paper annotation

In a first round of annotating papers with properties of human evaluations, we used the following paper-level properties, annotated using only information from the paper or supplementary material:

[2]Search performed in July 2022, so some TACL papers from later that year are not included.

1. How many systems were evaluated;
2. How many datasets were used;
3. Type of participant (e.g. MTurk);
4. How many unique participants;
5. Rough estimate of how many judgments;
6. Type of NLP task implemented by the system(s) evaluated (e.g. summarisation);
7. Input/output language(s) used (e.g. English).

During this first annotation, we manually filtered out papers only discussing human evaluation rather than including one (e.g., surveys of human evaluation), longitudinal studies, any that used highly specialised participants such as medical doctors, and any that we roughly estimated to be too expensive for us to repeat (threshold $2,000 in evaluator payments). This left 116 papers. For these papers, Table 3 in the appendix shows the counts[3] of the most common values for each property annotated. English was dominant as system language, used in over 90% of papers. The second most common language was Chinese, which was used in just under 10% of experiments. Language generation tasks were most common, with summarisation the most frequent task, followed by dialogue and MT.

About a third of papers did not specify type of participant. Among papers that did specify this, 60% used crowd-sourcing, with the vast majority of these being run on Mechanical Turk. It was generally difficult to find information about participants, with about half of papers not reporting the total number of participants. Very few papers included a clear description of the relationship between systems, data sets, items, and participants; number of judgments is therefore an estimate.

It became clear during high-level annotation that fewer than 5% of the 116 papers remaining after filtering were repeatable from publicly available information alone. Fundamental details like number and type of evaluators, instructions and training, and data evaluated are often omitted. Our next step was therefore to contact authors in the hope of obtaining the missing information. Lack of information about human evaluations has been commented on a number of times recently (van der Lee et al., 2019; Howcroft et al., 2020; Belz et al., 2020).

[3]Because some papers include multiple properties, for example, multiple languages in machine translation systems, some rows will not sum to 116.

Training or expertise	neither	only one	both
	11	13	4
Number of participants	small		not small
	14		14
Complexity	low	medium	high
	9	11	8

Table 1: Frequency of control-factor annotations.

3.2 More detailed annotation of experiments

In the next stage we carried out detailed annotation of evaluation properties preparatory to selecting a subset of such properties to control in our multi-lab multi-test study. We emailed the corresponding author (defaulting to first author) for each of the 116 papers to ask if they would support reproduction studies and, if they could provide more detailed information about their experiments.

The requested information included the user interface from the evaluation and the set of outputs shown to the evaluators (complete list see Appendix A.2). We received replies for just 39% of papers, even after sending reminders. Many of those who did reply were unable to provide the information needed. In the end, only 20 authors (20 papers containing 28 experiments) gave us enough information to progress the paper to the detailed annotation stage.[4] The most common reason for authors responding but being unable to provide information was that they had moved on from their (usually graduate student) position and files had not been kept. In some cases, authors from commercial research groups who were unable to provide information for business reasons. There were also eight papers where the authors responded initially, but the correspondence stalled.

Using the author-provided information together with paper, supplementary material and online resources, we annotated the 20 papers that progressed to this stage for the detailed properties of evaluations shown in Section A.4, annotated at the level of individual experiments (28), because at this more fine-grained annotation level, properties can differ between different experiments in the same paper.

One of the first three authors of the present paper annotated the 28 experiments with the detailed properties; the other two each checked half of the annotations. Any differences were discussed and

[4]One further author did provide sufficient information, but upon further analysis of the paper and the resources they sent, we decided that the evaluation experiment reported in it was too different from the other 20 papers; the systems detected change in language use over time.

resolved. To complete these annotations, we had to ask authors additional questions (usually in multiple rounds of questions and responses) for all experiments except two. In the end, for 8 of the 28 experiments we did not succeed in obtaining all the information needed for the above properties.

Note that the last two properties in Section A.4 (evaluation task complexity, interface complexity) have a different status from the others, in that they are secondary properties, subjectively assessed during annotation, rather than deriving from author-provided information. We found we tended to either agree on what their value should be, and when there was disagreement, values were adjacent. We used discussion rather than attempting to formalise rules to resolve disagreement, as it would seem an impossible task to exhaustively capture the latter.

Table 1, and Table 4 in the Appendix, show the frequency of the most common property values across the 28 experiments (here including unclear values). We found that most of the annotated properties have one or two values that are the most frequent by large margins. For example, assessments were *intrinsic* in 26 out of 28 experiments, *subjective* in 26 out of 28, and *absolute* in 20 out of 28. Only two experiments were *extrinsic* and *objective* evaluations, the other 26 were *intrinsic* and *subjective*. There was large variation in the number of participants, with a low of 2 and a high of 233. None of the experiments provided explicit training sessions for participants, and only one included a practice session. About three quarters of experiments provided instructions and/or criterion definitions.[5] Around half of the experiments used subjects with specialist expertise, which was usually linguistics or NLP.

3.3 Choosing properties to control for

The issues discussed in previous sections posed serious problems for selecting papers for a controlled study: we had only 20 fully annotated experiments; and we were left with very skewed distributions for many of the properties we had annotated, with many property combinations not occurring at all, or only occurring in one or two cases. Given the above issues it was clear that we were only going to be able to select a small set of properties to control for. We therefore whittled down the set of properties we had annotated to three that were both feasible and had a reasonable likelihood, based on existing work, of affecting reproducibility. For these, we created between two and three bins from the original value ranges, as follows:

1. ***Number of evaluators (small, not small)***: Experiments with 1–5 evaluators were assigned the *small* value, those with more than 5 evaluators the *not small* value.
2. ***Cognitive complexity of assessment performed by evaluators (low, medium, high)***: Experiments were assigned to one of the three possible values on the basis of the task complexity and interface complexity properties listed in Section A.4.
3. ***Training and/or expertise of evaluators (both, one, neither)***: Experiments that had both trained, and required specific expertise from, evaluators were assigned *both*; those that either trained evaluators or required expertise (but not both) were assigned *one*; the remainder were assigned *neither*.

Even for this much reduced set of control factors, we did not have enough experiments to cover all $2\times3\times3$ combinations of values, so we settled for a final set of 6 experiments, where there was an equal quantity of the pairwise combinations of the *Number of evaluators* and *Training/expertise* properties, as well as equal pairwise combinations of the *Number of evaluators* and *Complexity* properties.

4 Setting up Reproductions

Beginning the process of reproduction of the six experiments finally selected for reproduction (for common agreed approach to reproduction see Appendix A.5) necessarily involved delving into full implementational details for each of them. One particularly troubling finding has been the number of experimental flaws, errors and bugs we unearthed in the process. The more we dug into the properties of evaluation experiments that we needed in order to repeat an evaluation experiment, the more we uncovered flaws which made us question whether it made sense to repeat the experiment at all, in some cases because any conclusions drawn on the basis of the flawed experiments would be unsafe. Six specific issues are listed in Section A.6.[6] Note

[5] We cannot be precise because this information was in some cases not provided even after we interacted with authors.

[6] Note that we report these in anonymised form, because of the reputational risks involved. See also the Responsible Research Checklist included in the appendix.

Task	Num. Evaluators		Cognitive Complexity			Training and/or Experise		
	small	not small	low	medium	high	neither	either	both
Dialogue	1	0	0	1	0	0	1	0
Generation	6	5	4	5	2	4	5	2
Summarisation	3	1	2	1	1	1	3	0
Other	2	2	1	0	3	2	0	2

Table 2: Counts of control property values per NLP task for the 20 experiments (from 15 papers) where all properties were clear.

that only one of our six selected experiments had none of these issues. We are still discovering more.

The structure we designed for our original study is shown in the Appendix Section A.1, Figure 2.

5 Discussion

The reasons why we decided to abandon our original study design were as follows. One, we struggled to find enough papers that did not have (i) prohibitive barriers to reproduction, and/or (ii) unavailable information that would be needed for repeating experiments, and/or (iii) experimental flaws and errors. Two, no matter how much effort we put into obtaining full experimental details from authors, there still remained questions, albeit increasingly fine-grained, that we did not have the answer to, such as if the presentation order of evaluated items was randomised, or what instructions/training participants were given. In some cases, information about additional things that had been done, but could not be guessed from previously provided information, transpired coincidentally, necessitating further changes to experimental design.

A potential solution to not having enough papers at the end is selecting more papers at the start (more years, more events). However, given the inordinate amount of work we put into obtaining enough information from authors, simply tripling or quadrupling our initial pool of papers was not a viable solution. Similarly, there was little we were able to do about the reproduction barriers of excessive cost and highly specialised evaluators.

On the other hand, accepting to work from less than complete experimental information would have been problematic because information for different papers is incomplete in different ways, and we would not have been comparing like with like.

Correcting flaws and errors would similarly have introduced differences between original and reproduction studies, moreover different ones in different cases. In this case we would strictly speaking no longer have been conducting reproductions.

We considered designing new evaluations from scratch with the properties we wanted for our MLMT study. However, it would have been very difficult to ensure that newly created studies were somehow representative of the kind of studies that are actually being conducted in NLP.

We have now opted for a solution incorporating elements from most of the above, where we select a somewhat larger set of existing studies in a process similar to before, reduce the number of different values of factors we control for, and then *standardise and where necessary correct studies before reproduction*. Reproducibility is then measured between two new studies, rather than between them and the original study.

6 Conclusion

The track record of NLP as a field in recording information about human evaluation experiments is currently dire (Howcroft et al., 2020). We saw in the paper-level annotations (Appendix Table 3) that in 37 out of 116 papers the type of participant was unclear, in 59 the number of participants was unclear, and in 15 the number of judgements was unclear. Even after prolonged exchanges with authors during the experiment-level detailed annotation stage, very fundamental details were in some cases not obtainable: number of participants, details of training, instruction and practice items, whether participants were required to be native speakers, and even the set of outputs evaluated.

Our overall conclusion is that, on the basis of the unobtainability of information about experiments, barriers to reproduction and/or experimental flaws in our sample of 177 papers, only a small fraction of previous human evaluations in NLP can be repeated under the same conditions, hence that their reproducibility cannot be tested by repeating them. The way forward would appear to be to accept the overhead of detailed recording of experimental details, e.g. with HEDS (Shimorina and Belz, 2022), in combination with substantially increased standardisation in all aspects of experimental design.

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 Jackie Cheung.

Limitations

The small subset 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 phrases "human evaluation" and "participants." 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.

The small subset of conclusions based on our sample of experiments are limited by their sample size in terms of how representative they are 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 5, 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 (Appendix A.7).

References

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.

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.

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*, pages 52–61, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Joelle Pineau. 2020. The machine learning reproducibility checklist v2.0.

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.

Kirstie Whitaker. 2017. The MT Reproducibility Checklist. https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf.

A Appendix

A.1 Original study design

Figure 2 shows the original design of the multi-lab multi-test study.

A.2 Initial information requested from authors

Our initial email to authors asked if they would be able to provide the following information:

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.

A.3 Counts for high-level annotations

Table 3 shows counts for the first round of annotating paper-level properties.

A.4 Details of experiment-level annotation

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 (Belz et al., 2020):
   (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.

Structural design for a multi-lab, multi-test controlled study of experimental factors affecting reproducibility:

Round 1: Testing precision under repeatability conditions of measurement.

- Reproductions per experiment: 2 by two different labs;
- Conditions (experimental factors) to vary: evaluator cohort;
- If reproduction close enough, go to Round 2, else repeat Round 1 with improvements to experimental design, in terms of increased number of evaluators, and decreased cognitive complexity of evaluation task;
- For Round 1 repeats, if reproducibility is increased between reproduction studies (compared to each other, not the original study), proceed to Round 2, else stop.

Round 2: Testing reproducibility under varied conditions.

- Reproductions per experiment: 2 by two different labs;
- Conditions (experimental factors) to vary: evaluator cohort, and either number of evaluators *or* task complexity;
- If reproduction close enough, go to Round 3, else repeat Round 2 with improvements to experimental design, in terms of increased number of evaluators, and decreased cognitive complexity of evaluation task.
- For Round 2 repeats, if reproducibility is increased between reproduction studies (compared to each other, not the original study), proceed to Round 3, else stop.

Round 3: Testing reproducibility under increasingly varied conditions.

- Reproductions per experiment: 2 by two different labs;
- Conditions (experimental factors) to vary: evaluator cohort, number of evaluators *and* complexity.

Figure 2: Original design for the multi-lab, multi-test controlled study with a set of original human evaluation experiments with balanced experimental factors.

System language(s)	*English*	*Chinese*	*German*	*other*
	109	11	9	5
NLP Task	*summarisation*	*dialogue systems*	*machine translation*	*other*
	33	22	9	55
Number of systems	*1-5*	*6-7*	*> 7*	*unclear*
	89	14	13	0
Number of datasets	*1*	*2*	*> 3*	*unclear*
	83	25	8	0
Type of participant	*crowd (e.g., MTurk)*	*author/colleague/student*	*other*	*unclear*
	47	21	14	37
Number of unique participants	*< 5*	*5-20*	*> 20*	*unclear*
	27	19	11	59
Number of judgments	*< 100*	*100-1000*	*> 1000*	*unclear*
	1	34	66	15

Table 3: Frequency of the high-level experimental properties in the 116 papers, at the paper level. Some papers have multiple categorical properties therefore some rows will not sum to 116.

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.

A.5 Common Approach to Reproduction

In order to ensure comparability between studies, we agreed the following common-ground approach to carrying out reproduction studies:

1. Plan for repeating the original experiment identically, then apply to research ethics committee for approval.
2. If participants were paid during the original

Quality criteria names	*fluency*	*coherence*	*informativeness*	*other*
	10	5	3	54
System language(s)	*English*	*Chinese*	*German*	*other*
	26	3	2	0
NLP Task	*summarisation*	*question answering*	*explanation*	*other*
	6	3	3	16
Type of participant	*crowd*	*student*	*colleague*	*other*
	13	8	7	4
Intrinsic or extrinsic	*intrinsic*		*extrinsic*	
	26		2	
Absolute or relative	*absolute*		*relative*	
	20		8	
Objective or subjective	*objective*		*subjective*	
	2		26	
Num. of unique participants	*< 5*	*5–20*	*> 20*	*unclear*
	11	4	8	5
Num. of items evaluated	*< 200*	*200–1000*	*> 1000*	*unclear*
	9	10	7	2
Num. of participants per item	*< 4*	*4–9*	*> 9*	*varies*
	17	3	3	5
Num. of items per participant	*< 50*	*50–200*	*> 200*	*varies/unclear*
	5	5	7	11
Training given	*no*		*unclear*	
	24		4	
Instructions given	*yes*	*no*	*unclear*	
	8	15	5	
Criterion definitions given	*yes*	*no*	*n/a*	*unclear/mixed*
	17	3	4	4
Practice session held	*yes*	*no*	*unclear*	
	1	23	4	
Participant expertise type	*none*	*researcher*	*linguist*	*domain*
	16	9	2	1
Participants native speakers	*yes*	*no*	*of region*	*unknown*
	2	12	10	4

Table 4: Frequency of detailed experimental properties in set of 28 experiments.

experiment, determine pay in accordance with the common procedure for calculating fair pay (see appendix).

3. Complete HEDS datasheet.
4. Identify the following types of results reported in the original paper for the experiment:
   (a) Type I results: single numerical scores, e.g. mean quality rating, error count, etc.
   (b) Type II results: sets of numerical scores, e.g. set of Type I results.
   (c) Type III results: categorical labels attached to text spans of any length.
   (d) Qualitative conclusions/findings stated explicitly in the original paper.
5. Carry out the allocated experiment exactly as described in the HEDS sheet.
6. Report quantified reproducibility assessments for 8a–c as follows:
   (a) Type I results: Coefficient of variation (debiased for small samples).
   (b) Type II results: Pearson's r, Spearman's ρ.
   (c) Type III results: Multi-rater: Fleiss's κ; Multi-rater, multi-label: Krippendorff's α.
   (d) Conclusions/findings: Side-by-side summary of conclusions/findings that are / are not confirmed in the repeat experiment.

A.6 Issues, flaws and errors found

1. Mistakes in the reported figures for the human evaluation in the published paper, with the result that systems were reported as being better or worse that they actually were.
2. Reporting a total number of items in the paper which did not match the files that were sent.

3. Failure to randomise the order of items to be evaluated (when the stated intention was to randomise) due to wrongly applied randomisation.
4. Reporting that evaluators did equal numbers of assessments but it's clear from the files that they did very different numbers.
5. Ad-hoc attention checks (exact nature of which authors were unable to provide) applied to some but not all participants who if they failed the check were excluded from further contributing to the experiment, but whose already completed work was kept.
6. Biased methods of aggregating judgments (choosing a preferred participant rather than using some form of average).

On a more general note, ambiguities in the reporting can be an issue. Even when checked against the HEDS sheet, authors could feel like they have mentioned all experimental details that are asked for in HEDS, but often these are described at such a high level that there is still room for misinterpretation, which means that authors still need to confirm that their paper has been interpreted correctly. One solution for NLP authors could be to let a third party fill in the HEDS sheet and see where they get stuck, but this does add a further overhead.

A.7 ARR Responsible Research Checklist

A. **For every submission:**

A1. **Did you describe the limitations of your work?** Yes, e.g. we discuss the limitations from having a self-selecting subset of papers (where authors responded) available for analysis rather than a complete one.

A2. **Did you discuss any potential risks of your work?** The work analyses previously peer-reviewed and published human evaluation experiments, and while conventional risk considerations don't apply, we do mention the potential harm to individual authors from non-anonymously reporting experimental flaws and/or low reproducibility in their work.

A3. **Do the abstract and introduction summarise the paper's main claims?** Yes, abstract, introduction and conclusion summarise main aims and conclusions from the work.

B. **Did you use or create scientific artefacts?** No new data or computational resources were created.

C. **Did you run computational experiments?** No experiments were run.

D. **Did you use human annotators (e.g., crowdworkers) or research with human participants?** No human annotation or evaluations were carried out for this paper (other than by the authors).