The Human Evaluation Datasheet: A Template for Recording Details of Human Evaluation Experiments in NLP

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

Anya Belz
ADAPT Research Centre, DCU, Ireland
University of Aberdeen, UK
anya.belz@adaptcentre.ie

Abstract

This paper presents the Human Evaluation Datasheet (HEDS), a template for recording the details of individual human evaluation experiments in Natural Language Processing (NLP), and reports on first experience of researchers using HEDS sheets in practice. Originally taking inspiration from seminal papers by Bender and Friedman (2018), Mitchell et al. (2019), and Gebru et al. (2020), HEDS facilitates the recording of properties of human evaluations in sufficient detail, and with sufficient standardisation, to support comparability, meta-evaluation, and reproducibility assessments for human evaluations. These are crucial for scientifically principled evaluation, but the overhead of completing a detailed datasheet is substantial, and we discuss possible ways of addressing this and other issues observed in practice.

1 Introduction

Human evaluation plays a special role in NLP and NLG in particular as it is generally seen as the most reliable form of evaluation (Reiter and Belz, 2009; Novikova et al., 2017; Reiter, 2018). Comparability and reproducibility of evaluations (both human and automatic) are gaining in recognition and importance across NLP, as the field develops towards scientific maturity. For both reasons, it is of concern that there continues to be little consensus or standard practice across all aspects of human evaluation in NLP. Other efforts are aimed at standardisation of evaluation methods and quality criteria (van der Lee et al., 2021; Gehrmann et al., 2021). With the Human Evaluation Datasheet (HEDS) we wish to provide simply a way of recording properties of human evaluations in a standard hence comparable form (regardless of the level of standardisation of the deployed methods themselves).

HEDS is a template for recording properties of single human evaluation experiments. It uses multiple-choice questions where possible, for increased standardisation and automatic comparability. HEDS is designed to be generally applicable to human evaluations across NLP. It consists of 29 experiment-level questions plus 17 for each quality criterion, available as an online form which also contains explanations and guidance for completing it. The idea is that researchers use completed forms for preregistration of experiments and for archiving their details, to aid in comparability of evaluations across experiments, meta-evaluation of evaluation methods, and reproducibility of results.

Our intention is that HEDS should be suitable for all human evaluation experiments in NLP. Human evaluations in NLP typically get participants to assess system outputs or to interact with systems, but the HEDS sheet also accommodates what we call 'human-authored stand-ins' below, i.e. manually created 'system outputs' (e.g. in a wizard-of-oz scenario or when reference outputs are included in an evaluation) evaluated in a way that can at least in principle be used to evaluate actual system outputs.

The sheet is completed for a *single human evaluation experiment* by which we mean an experiment that evaluates a single set of directly comparable systems in a single experimental design, but may assess multiple quality criteria. This is the intended meaning when we refer to 'the evaluation experiment' in questions below.

2 Relationship to Existing Work

A first version of the datasheet (HEDS 1.0) was reported via a non-archival pre-print (Shimorina and Belz, 2021). In this paper, we present HEDS in its next revision (v2.0) alongside a summary of what we learnt from observing the datasheet being used in practice since its first publication. We focus discussion in this paper on Questions 4.1.1–4.2.3 (Sections 5 and 6) relating to quality criteria and their operationalisation: these caused some difficulty to users in practice, and were revised more substantively than other questions. The full

Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval), pages 54 - 75

datasheet is provided in the appendix for reference.

HEDS directly benefited from several papers and resources. Questions 2.1–2.5 relating to evaluated system, and 4.3.1–4.3.8 relating to response elicitation, are based on Howcroft et al. (2020), with some significant changes. Questions 4.1.1–4.2.3 relating to quality criteria, and some of the questions about system outputs, evaluators, and experimental design (3.1.1–3.2.3, 4.3.5, 4.3.6, 4.3.9–4.3.11) are based on Belz et al. (2020). HEDS was also informed by van der Lee et al. (2019, 2021) and by Gehrmann et al. (2021)'s[1] data card guide.

More generally, the original inspiration for creating a 'datasheet' for describing human evaluation experiments of course comes from seminal papers by Bender and Friedman (2018), Mitchell et al. (2019) and Gebru et al. (2020).

HEDS is related to other efforts in the NLP community related to reproducibility and ethics. Different NLP checklists have been introduced in recent years, e.g. the Reproducibility Checklist (Dodge et al., 2019) adopted by many conferences, and the ACL Rolling Review's Responsible NLP checklist.[2] These checklists mainly deal with the recreatability of computational experiments, details of used datasets and models, and risks and limitations of research studies and applications. The focus of HEDS is recording properties of human evaluation experiments, which are not covered by the above checklists.

3 HEDS Structure and Resources

The Human Evaluation Datasheet package consists of the following three resources:

1. The HEDS template: available at `https://forms.gle/MgWiKVu7i5UHeMNQ9`;
2. Description and completion guidance: this document;
3. Scripts for automatically converting between the HEDS online form and alternative Markdown and LaTeX template formats: available at `https://github.com/Shimorina/human-evaluation-datasheet`.

A collection of completed HEDS datasheets is also available at the HEDS GitHub repository.

The full HEDS sheet can be found in Appendix A. In its template form as well as in this paper, HEDS is divided into five sections, addressing topics and containing questions as follows:

1. Paper and Resources: HEDS Questions 1.1–1.3, listed in Appendix Section A.1;
2. Evaluated System: HEDS Questions 2.1–2.5, Section A.2;
3. Output Sample, Evaluators and Experimental Design: HEDS Questions 3.1.1–3.3.8, Section A.3;
4. Quality Criteria: HEDS Questions 4.1.1–4.3.11, listed in Section 6 of the paper and, for completeness also in Section A.4 in the Appendix—this section is completed separately for each quality criterion used as part of the same evaluation experiment;
5. Ethics: HEDS Questions 5.1–5.4, Section A.5.

Section A.1 records bibliographic information: link to the paper reporting the evaluation experiment, shared evaluation resources (e.g., a webpage, repository), contact author details.

Section A.2 describes information about outputs that are evaluated in the evaluation experiment and how they were produced. For example, it records the task performed by the system, types of system input and output, input and output language.

Section A.3 collects information about the evaluated sample (size, selection process, statistical power), the evaluators participating in the experiment, and experimental design (collection method, quality assurance, conditions for evaluators).

Section A.4 captures information about quality criteria assessed in the human evaluation experiment. We discuss this section in the main body of the paper in Section 6.

Section A.5 relates to ethical aspects of the evaluation: approval by ethics committees, and collection of personal and other sensitive data.

4 Insights from Use of HEDS 1.0 in Practice

HEDS 1.0 was used in the 2021 ReproGen Shared Task on Reproducibility of Evaluations in NLG[3] (Belz et al., 2021). All shared task participants, as well as the authors of the original papers up for

[1] `https://gem-benchmark.com/data_cards/guide`

[2] `https://aclrollingreview.org/responsibleNLPresearch/`

[3] `https://reprogen.github.io/2021/`

reproduction in Track A, completed the HEDS 1.0 form. Moreover, the HEDS 1.0 sheet was completed another three times outside of the shared task context.

General feedback from users was that the HEDS 1.0 sheet was for the most part straightforward to complete, but that completion still represented a considerable overhead. This is the general conundrum of initiatives such as HEDS: what is the incentive for researchers to complete the sheet when (i) it is merely good scientific hygiene rather than a component of the work without which it could not be carried out, and (ii) it mainly benefits follow on research rather than the paper it is completed for? Unless it is a requirement for submission to a particular event, or it is generally expected practice, the tendency may always be to avoid the overhead. To address this, we are in the process of preparing a shorter version of the sheet, with the aim of cutting the effort involved in half, for use in contexts where less detail is acceptable.

We also observed that there were a number of questions in HEDS 1.0 that users found difficult to complete for different reasons. Question 3.1.3 (*What is the statistical power of the sample size?*) posed difficulties partly because power calculations are a relatively new tool in NLP. We address this in this paper with additional information, and in the future by providing a new resource to support calculation.

A more fundamental issue was caused by how HEDS 1.0 captured information about quality criteria, and the use of duplicate subsections for multiple quality criteria. We believe that this was due largely to insufficient context, motivation and explanation being provided in the documentation and form about quality criteria and their properties, and we seek to provide the latter in version 2.0 (in this paper as well as in the form).

Using HEDS as part of the ReproGen shared task demonstrated the utility of having information about original studies and reproduction studies available in the same standard format: it meant it was straightforward for organisers to capture and analyse the similarities and differences between original and reproduction studies, e.g. to identify sources of variation in results for the results report (Belz et al., 2021). It also gave participants (authors of reproduction papers) a tool with which to verify whether their reproduction study of human evaluation was the same as the original study in all important respects, at a fine-grained level of detail.

5 Concepts Underlying Quality Criterion Questions 4.1.1–4.3.11 in HEDS

The overall aim of human evaluations in NLP is generally to assess some aspect of the quality of a system or component. Researchers use terms such as *Fluency* and *Informativeness* to refer to different aspects of quality. However, as discussed in detail by Howcroft et al. (2020) and Belz et al. (2020), just because two studies used the same term (e.g. Fluency) it does not mean they evaluated the same aspect of quality. In order to establish what was evaluated, we need to know the term and definition used, but also how it was 'operationalised,' i.e. what was presented to evaluators and how their assessments were recorded.

This is why HEDS, picking up from the two publications above, records properties relating to evaluation criteria and their operationalisation separately for each quality criterion. Because within the same experiment different quality criteria are often assessed in similar ways (e.g. using the same rating instrument), this can result in some repetition when completing a HEDS sheet, albeit unfortunately not in predictable ways.

Following Belz et al. (2020), properties relating to quality criteria and their operationalisation in HEDS fall into three groups: *quality criteria*, *evaluation mode*, and *experimental design*. A **quality criterion** is a criterion in terms of which the quality of system outputs is assessed, and is in itself entirely agnostic about how it is evaluated. **Evaluation modes** are properties that need to be specified to turn a quality criterion into an **evaluation measure** that can be implemented, and are orthogonal to quality criteria, i.e. any given quality criterion can be combined with any mode. **Experimental design** is the full specification of how to obtain a quantitative or qualitative *response value* for a given evaluation measure, yielding a fully specified **evaluation method**. In sum:

- Quality criterion + evaluation mode = evaluation measure;
- Evaluation measure + experimental design = evaluation method.

Each of the above concepts is covered by one or more questions in HEDS. Three HEDS questions capture properties of *quality criteria* in terms of (i) what type of quality is being assessed (Question 4.1.1); (ii) what aspect of the system output is

being assessed (Question 4.1.2); and (iii) whether system outputs are assessed in their own right or with reference to some system-internal or system-external frame of reference (Question 4.1.3).

Three questions capture *evaluation mode*: (i) subjective vs. objective (Question 4.2.1), (ii) absolute vs. relative (Question 4.2.2), and (iii) intrinsic vs. extrinsic (Question 4.2.3). *Experimental design* is covered by Questions 4.3.1–11 (operationalisation of quality criteria assessment), and Questions 3.* (other aspects).

We present, explain and discuss each of the above questions in the following section.

6 HEDS Questions about Properties of Quality Criteria and their Operationalisation

In this section, we present, verbatim, the questions referred to in the last section relating to properties of quality criteria and their operationalisation. All questions in this section need to be completed once for each quality criterion assessed in the single evaluation experiment that a HEDS sheet is being completed for. E.g. if an evaluation assesses *Fluency* and *Grammaticality*, then the questions below need to be filled in, separately, for each.

We refer below to 'output' as shorthand for that which is being assessed by evaluators. The latter is most often some form of language output assessed at different lengths (phrase, sentence, document), but it can also be a more complete form of system behaviour (e.g. language displayed along with audio and/or visual elements, on interfaces, etc.). It is in this more general sense that we intend the term 'output' to be understood in the present context.

6.1 Quality criterion properties (Questions 4.1.1–4.1.3)

Questions 4.1.1–4.1.3 capture the aspect of quality that is assessed by a given quality criterion, in terms of three orthogonal properties. They help determine whether or not the same aspect of quality is being evaluated in different evaluation experiments. The three properties characterise quality criteria in terms of (i) what type of quality is being assessed; (ii) what aspect of the system output is being assessed; and (iii) whether system outputs are assessed in their own right or with reference to some system-internal or system-external frame of reference. For full explanations see Belz et al. (2020).

Note that if NLP had a generally accepted standard set of quality criteria with common names, definitions and operationalisations, then most of the questions in this section could be replaced by a much smaller set capturing criterion name and operationalising techniques. The reason there are so many questions is precisely because we do not have such a standard nomenclature.

Question 4.1.1: What type of quality is assessed by the quality criterion?

Multiple-choice options (select one):

- ○ ***Correctness***: Select this option if it is possible to state, generally for all outputs, the conditions under which outputs are maximally correct (hence of maximal quality). E.g. for *Grammaticality*,[4] outputs are (maximally) correct if they contain no grammatical errors; for *Semantic Completeness*, outputs are correct if they express all the content in the input.
- ○ ***Goodness***: Select this option if, in contrast to correctness criteria, there is no single, general mechanism for deciding when outputs are maximally good, only for deciding for any two outputs which is better and which is worse. E.g. for *Fluency*, even if outputs contain no disfluencies, there may be other ways in which any given output could be more fluent.
- ○ ***Feature***: Choose this option if, in terms of property X captured by the criterion, outputs are not generally better if they are more X, but instead, depending on evaluation context, more X may be either better or worse. E.g. for *Specificity*, outputs can be more specific or less specific, but it's not the case that outputs are, in the general case, better when they are more specific.

Question 4.1.2: Which aspect of system outputs is assessed by the quality criterion?

Multiple-choice options (select one):

[4] We take all examples of quality criteria from published reports of evaluations, via the annotated database compiled by Howcroft et al. (2020).

- ***Form of output***: Choose this option if the criterion assesses the form of outputs alone, e.g. *Grammaticality* is only about the form, a sentence can be grammatical yet be wrong or nonsensical in terms of content.
- ***Content of output***: Select this option if the criterion assesses the content/meaning of the output alone, e.g. *Meaning Preservation* only assesses content; two sentences can be considered to have the same meaning, but differ in form.
- ***Both form and content of output***: Choose this option if the criterion assesses outputs as a whole, not just form or just content. E.g. *Coherence* is a property of outputs as a whole, either form or meaning can detract from it. Inherently extrinsic criteria such as *Usefulness* or *Task Completion* also fall in this category.

Question 4.1.3: Is each output assessed for quality in its own right, or with reference to a system-internal or external frame of reference?

Multiple-choice options (select one):

- ***Quality of output in its own right***: Select this option if output quality is assessed without referring to anything other than the output itself, i.e. no system-internal or external frame of reference. E.g. *Poeticness* is assessed by considering (just) the output and how poetic it is.
- ***Quality of output relative to the input***: Choose this option if output quality is assessed relative to the input. E.g. *Answerability* is the degree to which the output question can be answered from information in the input.
- ***Quality of output relative to a system-external frame of reference***: Choose this option if output quality is assessed with reference to system-external information, such as a knowledge base, a person's individual writing style, or the performance of an embedding system. E.g. *Factual Accuracy* assesses outputs relative to a source of real-world knowledge.

6.2 Evaluation mode properties (Questions 4.2.1–4.2.3)

Questions 4.2.1–4.2.3 record properties that are orthogonal to quality criteria (covered by questions in the preceding section), i.e. any given quality criterion can in principle be combined with any of the modes (although some combinations are more common than others).

Question 4.2.1: Does an individual assessment involve an objective or a subjective judgment?

Multiple-choice options (select one):

- ***Objective***: Choose this option if the evaluation uses objective assessment, e.g. any automatically counted or otherwise quantified measurements such as mouse-clicks, occurrences in text, etc. Repeated assessments of the same output with an objective-mode evaluation method always yield the same score/result.
- ***Subjective***: Choose this option in all other cases. Subjective assessments involve ratings, opinions and preferences by evaluators. Some criteria lend themselves more readily to subjective assessments, e.g. *Friendliness* of a conversational agent, but an objective measure e.g. based on lexical markers is also conceivable.

Question 4.2.2: Are outputs assessed in absolute or relative terms?

Multiple-choice options (select one):

- ***Absolute***: Select this option if evaluators are shown outputs from a single system during each individual assessment.
- ***Relative***: Choose this option if evaluators are shown outputs from multiple systems at the same time during assessments, typically ranking or preference-judging them.

Question 4.2.3: Is the evaluation intrinsic or extrinsic?

Multiple-choice options (select one):

- ***Intrinsic***: Choose this option if quality of outputs is assessed *without* considering their *effect* on something external to the system, e.g. the

performance of an embedding system or of a user at a task.

- ○ ***Extrinsic***: Choose this option if quality of outputs is assessed in terms of their *effect* on something external to the system such as the performance of an embedding system or of a user at a task.

6.3 Response elicitation (Questions 4.3.1–4.3.11)

The questions in this section concern response elicitation, by which we mean how the ratings or other measurements that represent assessments for the quality criterion in question are obtained, covering what is presented to evaluators, how they select response and via what type of tool, etc. The eleven questions (4.3.1–4.3.11) are based on the information annotated in the large scale survey of human evaluation methods in NLG by Howcroft et al. (2020).

Question 4.3.1: What do you call the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if criterion not named.

What to enter in the text box: the name you use to refer to the quality criterion in explanations and/or interfaces created for evaluators. Examples of quality criterion names include Fluency, Clarity, Meaning Preservation. If no name is used, state 'N/A'.

Question 4.3.2: What definition do you give for the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if no definition given.

What to enter in the text box: Copy and past the verbatim definition you give to evaluators to explain the quality criterion they're assessing. If you don't explicitly call it a definition, enter the nearest thing to a definition you give them. If you don't give any definition, state 'N/A'.

Question 4.3.3: Size of scale or other rating instrument (i.e. how many different possible values there are). Answer should be an integer or 'continuous' (if it's not possible to state how many possible responses there are). Enter 'N/A' if there is no rating instrument.

What to enter in the text box: The number of different response values for this quality criterion. E.g. for a 5-point Likert scale, the size to enter is 5. For two-way forced-choice preference judgments, it is 2; if there's also a no-preference option, enter 3. For a slider that is mapped to 100 different values for the purpose of recording assessments, the size to enter is 100. If no rating instrument is used (e.g. when evaluation gathers post-edits or qualitative feedback only), enter 'N/A'.

Question 4.3.4: List or range of possible values of the scale or other rating instrument. Enter 'N/A', if there is no rating instrument.

What to enter in the text box: list, or give the range of, the possible values of the rating instrument. The list or range should be of the size specified in Question 4.3.3. If there are too many to list, use a range. E.g. for two-way forced-choice preference judgments, the list entered might be *A better, B better*; if there's also a no-preference option, the list might be *A better, B better, neither*. For a slider that is mapped to 100 different values for the purpose of recording assessments, the range *1–100* might be entered. If no rating instrument is used (e.g. when evaluation gathers post-edits or qualitative feedback only), enter 'N/A'.

Question 4.3.5: How is the scale or other rating instrument presented to evaluators? If none match, select 'Other' and describe.

Multiple-choice options (select one):

- ○ ***Multiple-choice options***: choose this option if evaluators select exactly one of multiple options.

- ***Check-boxes***: choose this option if evaluators select any number of options from multiple given options.
- ***Slider***: choose this option if evaluators move a pointer on a slider scale to the position corresponding to their assessment.
- ***N/A (there is no rating instrument)***: choose this option if there is no rating instrument.
- ***Other (please specify)***: choose this option if there is a rating instrument, but none of the above adequately describe the way you present it to evaluators. Use the text box to describe the rating instrument and link to a screenshot.

Question 4.3.6: If there is no rating instrument, describe briefly what task the evaluators perform (e.g. ranking multiple outputs, finding information, playing a game, etc.), and what information is recorded. Enter 'N/A' if there is a rating instrument.

What to enter in the text box: If (and only if) there is no rating instrument, i.e. you entered 'N/A' for Questions 4.3.3–4.3.5, describe the task evaluators perform in this space. Otherwise, here enter 'N/A' if there *is* a rating instrument.

Question 4.3.7: What is the verbatim question, prompt or instruction given to evaluators (visible to them during each individual assessment)?

What to enter in the text box: Copy and paste the verbatim text that evaluators see during each assessment, that is intended to convey the evaluation task to them. E.g. *Which of these texts do you prefer?* Or *Make any corrections to this text that you think are necessary in order to improve it to the point where you would be happy to provide it to a client.*

Question 4.3.8: Form of response elicitation. If none match, select 'Other' and describe.

Multiple-choice options (select one):[5]

[5]Explanations adapted from Howcroft et al. (2020).

- ***(dis)agreement with quality statement***: Participants specify the degree to which they agree with a given quality statement by indicating their agreement on a rating instrument. The rating instrument is labelled with degrees of agreement and can additionally have numerical labels. E.g. *This text is fluent — 1=strongly disagree...5=strongly agree*.
- ***direct quality estimation***: Participants are asked to provide a rating using a rating instrument, which typically (but not always) mentions the quality criterion explicitly. E.g. *How fluent is this text? — 1=not at all fluent...5=very fluent*.
- ***relative quality estimation (including ranking)***: Participants evaluate two or more items in terms of which is better. E.g. *Rank these texts in terms of fluency*; *Which of these texts is more fluent?*; *Which of these items do you prefer?*.
- ***counting occurrences in text***: Evaluators are asked to count how many times some type of phenomenon occurs, e.g. the number of facts contained in the output that are inconsistent with the input.
- ***qualitative feedback (e.g. via comments entered in a text box)***: Typically, these are responses to open-ended questions in a survey or interview.
- ***evaluation through post-editing/annotation***: Choose this option if the evaluators' task consists of editing or inserting annotations in text. E.g. evaluators may perform error correction and edits are then automatically measured to yield a numerical score.
- ***output classification or labelling***: Choose this option if evaluators assign outputs to categories. E.g. *What is the overall sentiment of this piece of text? — Positive/neutral/negative.*
- ***user-text interaction measurements***: choose this option if participants in the evaluation experiment interact with a text in some way, and measurements are taken of their interaction. E.g. reading speed, eye movement tracking, comprehension questions, etc. Excludes situations where participants are given a task to solve and their performance is measured which comes under the next option.
- ***task performance measurements***: choose this option if participants in the evaluation experiment are given a task to perform, and measure-

ments are taken of their performance at the task. E.g. task is finding information, and task performance measurement is task completion speed and success rate.

- ***user-system interaction measurements***: choose this option if participants in the evaluation experiment interact with a system in some way, while measurements are taken of their interaction. E.g. duration of interaction, hyperlinks followed, number of likes, or completed sales.
- ***Other (please specify)***: Use the text box to describe the form of response elicitation used in assessing the quality criterion if it doesn't fall in any of the above categories.

Question 4.3.9: How are raw responses from participants aggregated or otherwise processed to obtain reported scores for this quality criterion? State if no scores reported.

What to enter in the text box: normally a set of separate assessments is collected from evaluators and is converted to the results as reported. Describe here the method(s) used in the conversion(s). E.g. macro-averages or micro-averages are computed from numerical scores to provide summary, per-system results.

Question 4.3.10: Method(s) used for determining effect size and significance of findings for this quality criterion.

What to enter in the text box: A list of methods used for calculating the effect size and significance of any results, both as reported in the paper given in Question 1.1, for this quality criterion. If none calculated, state 'None'.

Question 4.3.11: Has the inter-annotator and intra-annotator agreement between evaluators for this quality criterion been measured? If yes, what method was used, and what are the agreement scores?

What to enter in the text box: the methods used to compute, and results obtained from, any measures of inter-annotator and intra-annotator agreement obtained for the quality criterion.

7 Conclusion

In this paper we have presented the Human Evaluation Datasheet (HEDS), intended as a way of recording properties of human evaluations in NLP in a sufficiently standardised way to support comparability between evaluation experiments, meta-evaluation of evaluation methods, and reproducibility assessments of evaluation results.

We have reported insights from observing HEDS 1.0 being used in practice, and have described improvements we have made in response to these insights. In particular, we have provided additional context, motivation and explanation to the HEDS questions relating to evaluation criteria assessed in evaluation experiments and their operationalisation. Moreover, we are currently developing a shorter version of HEDS, a version with reduced effort for use in certain contexts.

We view HEDS as continuing to develop in response to feedback received and insights gathered through use in practice. We continue to welcome feedback on any aspect of HEDS, and hope the growing repository of completed sheets will prove useful for future comparisons, meta-evaluation and reproducibility assessments, as demonstrated in the ReproGen shared task.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

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, Anastasia Shimorina, Shubham Agarwal, and Ehud Reiter. 2021. The ReproGen shared task on reproducibility of human evaluations in NLG: Overview and results. In *The 14th International Conference on Natural Language Generation*.

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

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for datasets.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. 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.

David M. Howcroft and Verena Rieser. 2021. What happens if you treat ordinal ratings as interval data? human evaluations in NLP are even more underpowered than you think. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8932–8939, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hans Kamp and Uwe Reyle. 2013. *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*, volume 42. Springer Science & Business Media.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

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.

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

Ehud Reiter and Anya Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Anastasia Shimorina and Anya Belz. 2021. The human evaluation datasheet 1.0: A template for recording details of human evaluation experiments in NLP. ArXiv preprint arXiv:2103.09710v1.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.

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.

A Appendix: Full HEDS 2.0 Data Sheet

A.1 Questions about Paper and Supplementary Resources (Questions 1.1–1.3)

Questions 1.1–1.3 record bibliographic and related information. These are straightforward and don't warrant much in-depth explanation.

Question 1.1: Link to paper reporting the evaluation experiment. If the paper reports more than one experiment, state which experiment you're completing this sheet for. Or, if applicable, enter 'for preregistration.'

What to enter in the text box: a link to an online copy of the main reference for the human evaluation experiment, identifying which of the experiments the form is being completed for if there are several. If the experiment hasn't been run yet, and the form is being completed for the purpose of submitting it for preregistration, simply enter 'for preregistration'.

Question 1.2: Link to website providing resources used in the evaluation experiment (e.g. system outputs, evaluation tools, etc.). If there isn't one, enter 'N/A'.

What to enter in the text box: link(s) to any resources used in the evaluation experiment, such as system outputs, evaluation tools, etc. If there aren't any publicly shared resources (yet), enter 'N/A'.

Question 1.3: Name, affiliation and email address of person completing this sheet, and of contact author if different.

What to enter in the text box: names, affiliations and email addresses as appropriate.

A.2 System Questions 2.1–2.5

Questions 2.1–2.5 record information about the system(s) (or human-authored stand-ins) whose outputs are evaluated in the Evaluation experiment that this sheet is being completed for.

The input, output, and task questions in this section are closely interrelated: the value for one partially determines the others, as indicated for some combinations in Question 2.3.

Question 2.1: What type of input do the evaluated system(s) take? Select all that apply. If none match, select 'Other' and describe.

Describe the type of input, where input refers to the representations and/or data structures shared by all evaluated systems.

This question is about input type, regardless of number. E.g. if the input is a set of documents, you would still select *text: document* below.

Check-box options (select all that apply):

- ☐ ***raw/structured data***: numerical, symbolic, and other data, possibly structured into trees, graphs, graphical models, etc. May be the input e.g. to Referring Expression Generation (REG), end-to-end text generation, etc. NB: excludes linguistic structures.
- ☐ ***deep linguistic representation (DLR)***: any of a variety of deep, underspecified, semantic representations, such as abstract meaning representations (AMRs; Banarescu et al., 2013) or discourse representation structures (DRSs; Kamp and Reyle, 2013).
- ☐ ***shallow linguistic representation (SLR)***: any of a variety of shallow, syntactic representations, e.g. Universal Dependency (UD) structures; typically the input to surface realisation.
- ☐ ***text: subsentential unit of text***: a unit of text shorter than a sentence, e.g. Referring Expressions (REs), verb phrase, text fragment of any length; includes titles/headlines.
- ☐ ***text: sentence***: a single sentence (or set of sentences).
- ☐ ***text: multiple sentences***: a sequence of multiple sentences, without any document structure (or a set of such sequences).
- ☐ ***text: document***: a text with document structure, such as a title, paragraph breaks or sections, e.g. a set of news reports for summarisation.
- ☐ ***text: dialogue***: a dialogue of any length, excluding a single turn which would come under one of the other text types.

☐ ***text: other***: input is text but doesn't match any of the above *text:** categories.

☐ ***speech***: a recording of speech.

☐ ***visual***: an image or video.

☐ ***multi-modal***: catch-all value for any combination of data and/or linguistic representation and/or visual data etc.

☐ ***control feature***: a feature or parameter specifically present to control a property of the output text, e.g. positive stance, formality, author style.

☐ ***no input (human generation)***: human generation[6], therefore no system inputs.

☐ ***other (please specify)***: if input is none of the above, choose this option and describe it.

Question 2.2: What type of output do the evaluated system(s) generate? Select all that apply. If none match, select 'Other' and describe.

Describe the type of output, where output refers to the representations and/or data structures shared by all evaluated systems.

This question is about output type, regardless of number. E.g. if the output is a set of documents, you would still select *text: document* below.

Note that the options for outputs are the same as for inputs except that the *no input (human generation) option* is replaced with *human-generated 'outputs'*, and the *control feature* option is removed.

Check-box options (select all that apply):

☐ ***raw/structured data***: numerical, symbolic, and other data, possibly structured into trees, graphs, graphical models, etc. May be the input e.g. to Referring Expression Generation (REG), end-to-end text generation, etc. NB: excludes linguistic structures.

☐ ***deep linguistic representation (DLR)***: any of a variety of deep, underspecified, semantic representations, such as abstract meaning representations (AMRs; Banarescu et al., 2013) or discourse representation structures (DRSs; Kamp and Reyle, 2013).

[6]We use the term 'human generation' where the items being evaluated have been created manually, rather than generated by an automatic system.

☐ ***shallow linguistic representation (SLR)***: any of a variety of shallow, syntactic representations, e.g. Universal Dependency (UD) structures; typically the input to surface realisation.

☐ ***text: subsentential unit of text***: a unit of text shorter than a sentence, e.g. Referring Expressions (REs), verb phrase, text fragment of any length; includes titles/headlines.

☐ ***text: sentence***: a single sentence (or set of sentences).

☐ ***text: multiple sentences***: a sequence of multiple sentences, without any document structure (or a set of such sequences).

☐ ***text: document***: a text with document structure, such as a title, paragraph breaks or sections, e.g. a set of news reports for summarisation.

☐ ***text: dialogue***: a dialogue of any length, excluding a single turn which would come under one of the other text types.

☐ ***text: other***: select if output is text but doesn't match any of the above *text:** categories.

☐ ***speech***: a recording of speech.

☐ ***visual***: an image or video.

☐ ***multi-modal***: catch-all value for any combination of data and/or linguistic representation and/or visual data etc.

☐ ***human-generated 'outputs'***: manually created stand-ins exemplifying outputs.[6]

☐ ***other (please specify)***: if output is none of the above, choose this option and describe it.

Question 2.3: How would you describe the task that the evaluated system(s) perform in mapping the inputs in Q2.1 to the outputs in Q2.2? Occasionally, more than one of the options below may apply. If none match, select 'Other' and describe.

This field records the task performed by the system(s) being evaluated. This is independent of the application domain (financial reporting, weather forecasting, etc.), or the specific method (rule-based, neural, etc.) implemented in the system. We indicate mutual constraints between inputs, outputs and task for some of the options below.

Check-box options (select all that apply):

□ ***content selection/determination***: selecting the specific content that will be expressed in the generated text from a representation of possible content. This could be attribute selection for REG (without the surface realisation step). Note that the output here is not text.

□ ***content ordering/structuring***: assigning an order and/or structure to content to be included in generated text. Note that the output here is not text.

□ ***aggregation***: converting inputs (typically *deep linguistic representations* or *shallow linguistic representations*) in some way in order to reduce redundancy (e.g. representations for 'they like swimming', 'they like running' → representation for 'they like swimming and running').

□ ***referring expression generation***: generating *text* to refer to a given referent, typically represented in the input as a set of attributes or a linguistic representation.

□ ***lexicalisation***: associating (parts of) an input representation with specific lexical items to be used in their realisation.

□ ***deep generation***: one-step text generation from *raw/structured data* or *deep linguistic representations*. One-step means that no intermediate representations are passed from one independently run module to another.

□ ***surface realisation (SLR to text)***: one-step text generation from *shallow linguistic representations*. One-step means that no intermediate representations are passed from one independently run module to another.

□ ***feature-controlled text generation***: generation of text that varies along specific dimensions where the variation is controlled via *control feature*s specified as part of the input. Input is a non-textual representation (for feature-controlled text-to-text generation select the matching text-to-text task).

□ ***data-to-text generation***: generation from *raw/structured data* which may or may not include some amount of content selection as part of the generation process. Output is likely to be *text:** or *multi-modal*.

□ ***dialogue turn generation***: generating a dialogue turn (can be a greeting or closing) from a representation of dialogue state and/or last turn(s), etc.

□ ***question generation***: generation of questions from given input text and/or knowledge base such that the question can be answered from the input.

□ ***question answering***: input is a question plus optionally a set of reference texts and/or knowledge base, and the output is the answer to the question.

□ ***paraphrasing/lossless simplification***: text-to-text generation where the aim is to preserve the meaning of the input while changing its wording. This can include the aim of changing the text on a given dimension, e.g. making it simpler, changing its stance or sentiment, etc., which may be controllable via input features. Note that this task type includes meaning-preserving text simplification (non-meaning preserving simplification comes under *compression/lossy simplification* below).

□ ***compression/lossy simplification***: text-to-text generation that has the aim to generate a shorter, or shorter and simpler, version of the input text. This will normally affect meaning to some extent, but as a side effect, rather than the primary aim, as is the case in *summarisation*.

□ ***machine translation***: translating text in a source language to text in a target language while maximally preserving the meaning.

□ ***summarisation (text-to-text)***: output is an extractive or abstractive summary of the important/relevant/salient content of the input document(s).

□ ***end-to-end text generation***: use this option if the single system task corresponds to more than one of tasks above, implemented either as separate modules pipelined together, or as one-step generation, other than *deep generation* and *surface realisation*.

□ ***image/video description***: input includes *visual*, and the output describes it in some way.

□ ***post-editing/correction***: system edits and/or corrects the input text (typically itself the textual output from another system) to yield an improved version of the text.

□ ***other (please specify)***: if task is none of the above, choose this option and describe it.

Question 2.4: Input Language(s), or 'N/A'.

This field records the language(s) of the inputs accepted by the system(s) being evaluated.

What to enter in the text box: any language name(s) that apply, mapped to standardised full language names in ISO 639-1[7]. E.g. English, Herero, Hindi. If no language is accepted as (part of) the input, enter 'N/A'.

Question 2.5: Output Language(s), or 'N/A'.

This field records the language(s) of the outputs generated by the system(s) being evaluated.

What to enter in the text box: any language name(s) that apply, mapped to standardised full language names in ISO 639-1 (2019)[7]. E.g. English, Herero, Hindi. If no language is generated, enter 'N/A'.

A.3 Questions about Output Sample, Evaluators, Experimental Design

A.3.1 Sample of system outputs (or human-authored stand-ins) evaluated (Questions 3.1.1–3.1.3)

Questions 3.1.1–3.1.3 record information about the size of the sample of outputs (or human-authored stand-ins) evaluated per system, how the sample was selected, and what its statistical power is.

Question 3.1.1: How many system outputs (or other evaluation items) are evaluated per system in the evaluation experiment? Answer should be an integer.

What to enter in the text box: The number of system outputs (or other evaluation items) that are evaluated per system by at least one evaluator in the experiment, as an integer.

Question 3.1.2: How are system outputs (or other evaluation items) selected for inclusion in the evaluation experiment? If none match, select 'Other' and describe.

Multiple-choice options (select one):

- ○ ***by an automatic random process from a larger set***: outputs were selected for inclusion in the experiment by a script using a pseudo-random number generator; don't use this option if the script selects every nth output (which is not random).
- ○ ***by an automatic random process but using stratified sampling over given properties***: use this option if selection was by a random script as above, but with added constraints ensuring that the sample is representative of the set of outputs it was selected from, in terms of given properties, such as sentence length, positive/negative stance, etc.
- ○ ***by manual, arbitrary selection***: output sample was selected by hand, or automatically from a manually compiled list, without a specific selection criterion.
- ○ ***by manual selection aimed at achieving balance or variety relative to given properties***: selection by hand as above, but with specific selection criteria, e.g. same number of outputs from each time period.
- ○ ***Other (please specify)***: if selection method is none of the above, choose this option and describe it.

Question 3.1.3: What is the statistical power of the sample size?

What to enter in the text box: The results of a statistical power calculation on the output sample: provide numerical results and a link to the script used (or another way of identifying the script). See, e.g., Card et al. (2020); Howcroft and Rieser (2021).

A.3.2 Evaluators (Questions 3.2.1–3.2.5)

Questions 3.2.1–3.2.5 record information about the evaluators participating in the experiment.

Question 3.2.1: How many evaluators are there in this experiment? Answer should be an integer.

What to enter in the text box: the total number of evaluators participating in the experiment, as an integer.

[7] https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

Question 3.2.2: What kind of evaluators are in this experiment? Select all that apply. If none match, select 'Other' and describe. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

- ☐ ***experts***: participants are considered domain experts, e.g. meteorologists evaluating a weather forecast generator, or nurses evaluating an ICU report generator.
- ☐ ***non-experts***: participants are not domain experts.
- ☐ ***paid (including non-monetary compensation such as course credits)***: participants were given some form of compensation for their participation, including vouchers, course credits, and reimbursement for travel unless based on receipts.
- ☐ ***not paid***: participants were not given compensation of any kind.
- ☐ ***previously known to authors***: (one of the) researchers running the experiment knew some or all of the participants before recruiting them for the experiment.
- ☐ ***not previously known to authors***: none of the researchers running the experiment knew any of the participants before recruiting them for the experiment.
- ☐ ***evaluators include one or more of the authors***: one or more researchers running the experiment was among the participants.
- ☐ ***evaluators do not include any of the authors***: none of the researchers running the experiment were among the participants.
- ☐ ***Other*** (fewer than 4 of the above apply): we believe you should be able to tick 4 options of the above. If that's not the case, use this box to explain.

Question 3.2.3: How are evaluators recruited?

What to enter in the text box: Please explain how your evaluators are recruited. Do you send emails to a given list? Do you post invitations on social media? Posters on university walls? Were there any gatekeepers involved? What are the exclusion/inclusion criteria?

Question 3.2.4: What training and/or practice are evaluators given before starting on the evaluation itself?

What to enter in the text box: Use this space to describe any training evaluators were given as part of the experiment to prepare them for the evaluation task, including any practice evaluations they did. This includes any introductory explanations they're given, e.g. on the start page of an online evaluation tool.

Question 3.2.5: What other characteristics do the evaluators have, known either because these were qualifying criteria, or from information gathered as part of the evaluation?

What to enter in the text box: Use this space to list any characteristics not covered in previous questions that the evaluators are known to have, either because evaluators were selected on the basis of a characteristic, or because information about a characteristic was collected as part of the evaluation. This might include geographic location of IP address, educational level, or demographic information such as gender, age, etc. Where characteristics differ among evaluators (e.g. gender, age, location etc.), also give numbers for each subgroup.

A.3.3 Experimental Design Questions 3.3.1–3.3.8

Questions 3.3.1–3.3.8 record information about the experimental design of the evaluation experiment.

Question 3.3.1: Has the experimental design been preregistered? If yes, on which registry?

What to enter in the text box: State 'Yes' or 'No'; if 'Yes' also give the name of the registry and a link to the registration page for the experiment.

Question 3.3.2: How are responses collected? E.g. paper forms, online survey tool, etc.

What to enter in the text box: Use this space to describe how you collected responses, e.g. paper forms, Google forms, SurveyMonkey, Mechanical Turk, CrowdFlower, audio/video recording, etc.

Question 3.3.3: What quality assurance methods are used? Select all that apply. If none match, select 'Other' and describe. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

- ☐ ***evaluators are required to be native speakers of the language they evaluate***: mechanisms are in place to ensure all participants are native speakers of the language they evaluate.
- ☐ ***automatic quality checking methods are used during/post evaluation***: evaluations are checked for quality by automatic scripts during or after evaluations, e.g. evaluators are given known bad/good outputs to check they're given bad/good scores on MTurk.
- ☐ ***manual quality checking methods are used during/post evaluation***: evaluations are checked for quality by a manual process during or after evaluations, e.g. scores assigned by evaluators are monitored by researchers conducting the experiment.
- ☐ ***evaluators are excluded if they fail quality checks (often or badly enough)***: there are conditions under which evaluations produced by participants are not included in the final results due to quality issues.
- ☐ ***some evaluations are excluded because of failed quality checks***: there are conditions under which some (but not all) of the evaluations produced by some participants are not included in the final results due to quality issues.
- ☐ ***none of the above***: tick this box if none of the above apply.
- ☐ ***Other (please specify)***: use this box to describe any other quality assurance methods used during or after evaluations, and to provide additional details for any of the options selected above.

Question 3.3.4: What do evaluators see when carrying out evaluations? Link to screenshot(s) and/or describe the evaluation interface(s).

What to enter in the text box: Use this space to describe the interface, paper form, etc. that evaluators see when they carry out the evaluation. Link to a screenshot/copy if possible. If there is a separate introductory interface/page, include it under Question 3.2.4.

3.3.5: How free are evaluators regarding when and how quickly to carry out evaluations? Select all that apply. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

- ☐ ***evaluators have to complete each individual assessment within a set time***: evaluators are timed while carrying out each assessment and cannot complete the assessment once time has run out.
- ☐ ***evaluators have to complete the whole evaluation in one sitting***: partial progress cannot be saved and the evaluation returned to on a later occasion.
- ☐ ***neither of the above***: Choose this option if neither of the above are the case in the experiment.
- ☐ ***Other (please specify)***: Use this space to describe any other way in which time taken or number of sessions used by evaluators is controlled in the experiment, and to provide additional details for any of the options selected above.

3.3.6: Are evaluators told they can ask questions about the evaluation and/or provide feedback? Select all that apply. In all cases, provide details in the text box under 'Other'.

Check-box options (select all that apply):

- ☐ ***evaluators are told they can ask any questions during/after receiving initial training/instructions, and before the start of the evaluation***: evaluators are told explicitly that they can ask questions about the evaluation experiment *before* starting on their assessments, either during or after training.
- ☐ ***evaluators are told they can ask any questions during the evaluation***: evaluators are told explicitly that they can ask questions about the evaluation experiment *during* their assessments.
- ☐ ***evaluators are asked for feedback and/or comments after the evaluation, e.g. via an exit questionnaire or a comment box***: evaluators are explicitly asked to provide feedback and/or comments about the experiment *after* their assessments, either verbally or in written form.
- ☐ ***None of the above***: Choose this option if none of the above are the case in the experiment.
- ☐ ***Other (please specify)***: use this space to describe any other ways you provide for evaluators to ask questions or provide feedback.

3.3.7: What are the experimental conditions in which evaluators carry out the evaluations? If none match, select 'Other' and describe.

Multiple-choice options (select one):

- ○ ***evaluation carried out by evaluators at a place of their own choosing, e.g. online, using a paper form, etc.***: evaluators are given access to the tool or form specified in Question 3.3.2, and subsequently choose where to carry out their evaluations.
- ○ ***evaluation carried out in a lab, and conditions are the same for each evaluator***: evaluations are carried out in a lab, and conditions in which evaluations are carried out *are* controlled to be the same, i.e. the different evaluators all carry out the evaluations in identical conditions of quietness, same type of computer, same room, etc. Note we're not after very fine-grained differences here, such as time of day or temperature, but the line is difficult to draw, so some judgment is involved here.
- ○ ***evaluation carried out in a lab, and conditions vary for different evaluators***: choose this option if evaluations are carried out in a lab, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.
- ○ ***evaluation carried out in a real-life situation, and conditions are the same for each evaluator***: evaluations are carried out in a real-life situation, i.e. one that would occur whether or not the evaluation was carried out (e.g. evaluating a dialogue system deployed in a live chat function on a website), and conditions in which evaluations are carried out *are* controlled to be the same.
- ○ ***evaluation carried out in a real-life situation, and conditions vary for different evaluators***: choose this option if evaluations are carried out in a real-life situation, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.
- ○ ***evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions are the same for each evaluator***: evaluations are carried out outside of the lab, in a situation intentionally similar to a real-life situation (but not actually a real-life situation), e.g. user-testing a navigation system where the destination is part of the evaluation design, rather than chosen by the user. Conditions in which evaluations are carried out *are* controlled to be the same.
- ○ ***evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions vary for different evaluators***: choose this option if evaluations are carried out outside of the lab, in a situation intentionally similar to a real-life situation, but the preceding option does not apply, i.e. conditions in which evaluations are carried out are *not* controlled to be the same.
- ○ ***Other (please specify)***: Use this space to provide additional, or alternative, information

about the conditions in which evaluators carry out assessments, not covered by the options above.

3.3.8: Unless the evaluation is carried out at a place of the evaluators' own choosing, briefly describe the (range of different) conditions in which evaluators carry out the evaluations.

What to enter in the text box: use this space to describe the variations in the conditions in which evaluators carry out the evaluation, for both situations where those variations are controlled, and situations where they are not controlled.

A.4 Quality Criterion *n* – Definition and Operationalisation

Questions in this section collect information about the nth quality criterion assessed in the single human evaluation experiment that this sheet is being completed for.

A.4.1 Quality criterion properties (Questions 4.1.1–4.1.3)

Questions 4.1.1–4.1.3 capture the aspect of quality that is assessed by a given quality criterion in terms of three orthogonal properties. They help determine whether or not the same aspect of quality is being evaluated in different evaluation experiments. The three properties characterise quality criteria in terms of (i) what type of quality is being assessed; (ii) what aspect of the system output is being assessed; and (iii) whether system outputs are assessed in their own right or with reference to some system-internal or system-external frame of reference. For full explanations see Belz et al. (2020).

Question 4.1.1: What type of quality is assessed by the quality criterion?

Multiple-choice options (select one):

- ○ ***Correctness***: Select this option if it is possible to state, generally for all outputs, the conditions under which outputs are maximally correct (hence of maximal quality). E.g. for *Grammaticality*,[8] outputs are (maximally) correct if they contain no grammatical errors; for *Semantic Completeness*, outputs are correct if they express all the content in the input.
- ○ ***Goodness***: Select this option if, in contrast to correctness criteria, there is no single, general mechanism for deciding when outputs are maximally good, only for deciding for any two outputs which is better and which is worse. E.g. for *Fluency*, even if outputs contain no disfluencies, there may be other ways in which any given output could be more fluent.
- ○ ***Feature***: Choose this option if, in terms of property X captured by the criterion, outputs are not generally better if they are more X, but instead, depending on evaluation context, more X may be either better or worse. E.g. for *Specificity*, outputs can be more specific or less specific, but it's not the case that outputs are, in the general case, better when they are more specific.

Question 4.1.2: Which aspect of system outputs is assessed by the quality criterion?

Multiple-choice options (select one):

- ○ ***Form of output***: Choose this option if the criterion assesses the form of outputs alone, e.g. *Grammaticality* is only about the form, a sentence can be grammatical yet be wrong or nonsensical in terms of content.
- ○ ***Content of output***: Select this option if the criterion assesses the content/meaning of the output alone, e.g. *Meaning Preservation* only assesses content; two sentences can be considered to have the same meaning, but differ in form.
- ○ ***Both form and content of output***: Choose this option if the criterion assesses outputs as a whole, not just form or just content. E.g. *Coherence* is a property of outputs as a whole, either form or meaning can detract from it. Inherently extrinsic criteria such as *Usefulness* or *Task Completion* also fall in this category.

[8] We take all examples of quality criteria from published reports of evaluations, via the annotated database compiled by Howcroft et al. (2020).

Question 4.1.3: Is each output assessed for quality in its own right, or with reference to a system-internal or external frame of reference?

Multiple-choice options (select one):

- ○ ***Quality of output in its own right***: Select this option if output quality is assessed without referring to anything other than the output itself, i.e. no system-internal or external frame of reference. E.g. *Poeticness* is assessed by considering (just) the output and how poetic it is.
- ○ ***Quality of output relative to the input***: Choose this option if output quality is assessed relative to the input. E.g. *Answerability* is the degree to which the output question can be answered from information in the input.
- ○ ***Quality of output relative to a system-external frame of reference***: Choose this option if output quality is assessed with reference to system-external information, such as a knowledge base, a person's individual writing style, or the performance of an embedding system. E.g. *Factual Accuracy* assesses outputs relative to a source of real-world knowledge.

A.4.2 Evaluation mode properties (Questions 4.2.1–4.2.3)

Questions 4.2.1–4.2.3 record properties that are orthogonal to quality criteria (covered by questions in the preceding section), i.e. any given quality criterion can in principle be combined with any of the modes (although some combinations are more common than others).

Question 4.2.1: Does an individual assessment involve an objective or a subjective judgment?

Multiple-choice options (select one):

- ○ ***Objective***: Choose this option if the evaluation uses objective assessment, e.g. any automatically counted or otherwise quantified measurements such as mouse-clicks, occurrences in text, etc. Repeated assessments of the same output with an objective-mode evaluation method always yield the same score/result.
- ○ ***Subjective***: Choose this option in all other cases. Subjective assessments involve ratings, opinions and preferences by evaluators. Some criteria lend themselves more readily to subjective assessments, e.g. *Friendliness* of a conversational agent, but an objective measure e.g. based on lexical markers is also conceivable.

Question 4.2.2: Are outputs assessed in absolute or relative terms?

Multiple-choice options (select one):

- ○ ***Absolute***: Select this option if evaluators are shown outputs from a single system during each individual assessment.
- ○ ***Relative***: Choose this option if evaluators are shown outputs from multiple systems at the same time during assessments, typically ranking or preference-judging them.

Question 4.2.3: Is the evaluation intrinsic or extrinsic?

Multiple-choice options (select one):

- ○ ***Intrinsic***: Choose this option if quality of outputs is assessed *without* considering their *effect* on something external to the system, e.g. the performance of an embedding system or of a user at a task.
- ○ ***Extrinsic***: Choose this option if quality of outputs is assessed in terms of their *effect* on something external to the system such as the performance of an embedding system or of a user at a task.

A.4.3 Response elicitation (Questions 4.3.1–4.3.11)

The questions in this section concern response elicitation, by which we mean how the ratings or other measurements that represent assessments for the quality criterion in question are obtained, covering what is presented to evaluators, how they select response and via what type of tool, etc. The eleven questions (4.3.1–4.3.11) are based on the information annotated in the large scale survey of human evaluation methods in NLG by Howcroft et al. (2020).

Question 4.3.1: What do you call the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if criterion not named.

What to enter in the text box: the name you use to refer to the quality criterion in explanations and/or interfaces created for evaluators. Examples of quality criterion names include Fluency, Clarity, Meaning Preservation. If no name is used, state 'N/A'.

Question 4.3.2: What definition do you give for the quality criterion in explanations/interfaces to evaluators? Enter 'N/A' if no definition given.

What to enter in the text box: Copy and past the verbatim definition you give to evaluators to explain the quality criterion they're assessing. If you don't explicitly call it a definition, enter the nearest thing to a definition you give them. If you don't give any definition, state 'N/A'.

Question 4.3.3: Size of scale or other rating instrument (i.e. how many different possible values there are). Answer should be an integer or 'continuous' (if it's not possible to state how many possible responses there are). Enter 'N/A' if there is no rating instrument.

What to enter in the text box: The number of different response values for this quality criterion. E.g. for a 5-point Likert scale, the size to enter is 5. For two-way forced-choice preference judgments, it is 2; if there's also a no-preference option, enter 3. For a slider that is mapped to 100 different values for the purpose of recording assessments, the size to enter is 100. If no rating instrument is used (e.g. when evaluation gathers post-edits or qualitative feedback only), enter 'N/A'.

Question 4.3.4: List or range of possible values of the scale or other rating instrument. Enter 'N/A', if there is no rating instrument.

What to enter in the text box: list, or give the range of, the possible values of the rating instrument. The list or range should be of the size specified in Question 4.3.3. If there are too many to list, use a range. E.g. for two-way forced-choice preference judgments, the list entered might be *A better, B better*; if there's also a no-preference option, the list might be *A better, B better, neither*. For a slider that is mapped to 100 different values for the purpose of recording assessments, the range *1–100* might be entered. If no rating instrument is used (e.g. when evaluation gathers post-edits or qualitative feedback only), enter 'N/A'.

Question 4.3.5: How is the scale or other rating instrument presented to evaluators? If none match, select 'Other' and describe.

Multiple-choice options (select one):

- ***Multiple-choice options***: choose this option if evaluators select exactly one of multiple options.
- ***Check-boxes***: choose this option if evaluators select any number of options from multiple given options.
- ***Slider***: choose this option if evaluators move a pointer on a slider scale to the position corresponding to their assessment.
- ***N/A (there is no rating instrument)***: choose this option if there is no rating instrument.
- ***Other (please specify)***: choose this option if there is a rating instrument, but none of the above adequately describe the way you present it to evaluators. Use the text box to describe the rating instrument and link to a screenshot.

Question 4.3.6: If there is no rating instrument, describe briefly what task the evaluators perform (e.g. ranking multiple outputs, finding information, playing a game, etc.), and what information is recorded. Enter 'N/A' if there is a rating instrument.

What to enter in the text box: If (and only if) there is no rating instrument, i.e. you entered 'N/A' for Questions 4.3.3–4.3.5, describe the task evaluators

perform in this space. Otherwise, here enter 'N/A' if there *is* a rating instrument.

Question 4.3.7: What is the verbatim question, prompt or instruction given to evaluators (visible to them during each individual assessment)?

What to enter in the text box: Copy and paste the verbatim text that evaluators see during each assessment, that is intended to convey the evaluation task to them. E.g. *Which of these texts do you prefer?* Or *Make any corrections to this text that you think are necessary in order to improve it to the point where you would be happy to provide it to a client.*

Question 4.3.8: Form of response elicitation. If none match, select 'Other' and describe.

Multiple-choice options (select one):[9]

- ○ ***(dis)agreement with quality statement***: Participants specify the degree to which they agree with a given quality statement by indicating their agreement on a rating instrument. The rating instrument is labelled with degrees of agreement and can additionally have numerical labels. E.g. *This text is fluent — 1=strongly disagree...5=strongly agree.*
- ○ ***direct quality estimation***: Participants are asked to provide a rating using a rating instrument, which typically (but not always) mentions the quality criterion explicitly. E.g. *How fluent is this text? — 1=not at all fluent...5=very fluent.*
- ○ ***relative quality estimation (including ranking)***: Participants evaluate two or more items in terms of which is better. E.g. *Rank these texts in terms of fluency*; *Which of these texts is more fluent?*; *Which of these items do you prefer?*.
- ○ ***counting occurrences in text***: Evaluators are asked to count how many times some type of phenomenon occurs, e.g. the number of facts contained in the output that are inconsistent with the input.
- ○ ***qualitative feedback (e.g. via comments entered in a text box)***: Typically, these are responses to open-ended questions in a survey or interview.
- ○ ***evaluation through post-editing/annotation***: Choose this option if the evaluators' task consists of editing or inserting annotations in text. E.g. evaluators may perform error correction and edits are then automatically measured to yield a numerical score.
- ○ ***output classification or labelling***: Choose this option if evaluators assign outputs to categories. E.g. *What is the overall sentiment of this piece of text? — Positive/neutral/negative.*
- ○ ***user-text interaction measurements***: choose this option if participants in the evaluation experiment interact with a text in some way, and measurements are taken of their interaction. E.g. reading speed, eye movement tracking, comprehension questions, etc. Excludes situations where participants are given a task to solve and their performance is measured which comes under the next option.
- ○ ***task performance measurements***: choose this option if participants in the evaluation experiment are given a task to perform, and measurements are taken of their performance at the task. E.g. task is finding information, and task performance measurement is task completion speed and success rate.
- ○ ***user-system interaction measurements***: choose this option if participants in the evaluation experiment interact with a system in some way, while measurements are taken of their interaction. E.g. duration of interaction, hyperlinks followed, number of likes, or completed sales.
- ○ ***Other (please specify)***: Use the text box to describe the form of response elicitation used in assessing the quality criterion if it doesn't fall in any of the above categories.

Question 4.3.9: How are raw responses from participants aggregated or otherwise processed to obtain reported scores for this quality criterion? State if no scores reported.

What to enter in the text box: normally a set of separate assessments is collected from evaluators

[9] Explanations adapted from Howcroft et al. (2020).

and is converted to the results as reported. Describe here the method(s) used in the conversion(s). E.g. macro-averages or micro-averages are computed from numerical scores to provide summary, per-system results.

Question 4.3.10: Method(s) used for determining effect size and significance of findings for this quality criterion.

What to enter in the text box: A list of methods used for calculating the effect size and significance of any results, both as reported in the paper given in Question 1.1, for this quality criterion. If none calculated, state 'None'.

Question 4.3.11: Has the inter-annotator and intra-annotator agreement between evaluators for this quality criterion been measured? If yes, what method was used, and what are the agreement scores?

What to enter in the text box: the methods used to compute, and results obtained from, any measures of inter-annotator and intra-annotator agreement obtained for the quality criterion.

A.5 Ethics Questions (Questions 5.1-5.4)

The questions in this section relate to ethical aspects of the evaluation. Information can be entered in the text box provided, and/or by linking to a source where complete information can be found.

Question 5.1: Has the evaluation experiment this sheet is being completed for, or the larger study it is part of, been approved by a research ethics committee? If yes, which research ethics committee?

What to enter in the text box: Typically, research organisations, universities and other higher-education institutions require some form ethical approval before experiments involving human participants, however innocuous, are permitted to proceed. Please provide here the name of the body that approved the experiment, or state 'No' if approval has not (yet) been obtained.

Question 5.2: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain personal data (as defined in GDPR Art. 4, §1: https://gdpr.eu/article-4-definitions/)? If yes, describe data and state how addressed.

What to enter in the text box: State 'No' if no personal data as defined by GDPR was recorded or collected, otherwise explain how conformity with GDPR requirements such as privacy and security was ensured, e.g. by linking to the (successful) application for ethics approval from Question 5.1.

Question 5.3: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain special category information (as defined in GDPR Art. 9, §1: https://gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited/)? If yes, describe data and state how addressed.

What to enter in the text box: State 'No' if no special-category data as defined by GDPR was recorded or collected, otherwise explain how conformity with GDPR requirements relating to special-category data was ensured, e.g. by linking to the (successful) application for ethics approval from Question 5.1.

Question 5.4: Have any impact assessments been carried out for the evaluation experiment, and/or any data collected/evaluated in connection with it? If yes, summarise approach(es) and outcomes.

What to enter in the text box: Use this box to describe any *ex ante* or *ex post* impact assessments that have been carried out in relation to the evaluation experiment, such that the assessment plan and process, as well as the outcomes, were captured in written form. Link to documents if possible. Types of impact assessment include data protection

impact assessments, e.g. under GDPR.[10] Environmental and social impact assessment frameworks are also available.

[10] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/