

# HEDS 3.0: The Human Evaluation Data Sheet Version 3.0

**Anya Belz and Craig Thomson**  
ADAPT, Dublin City University  
Dublin, Ireland  
{anya.belz, craig.thomson}@dcu.ie

## Abstract

This paper presents the Human Evaluation Datasheet (HEDS) Version 3.0. This update is the result of our experience using HEDS in the context of numerous recent human evaluation experiments, including reproduction studies, and of feedback collected from other researchers. HEDS 3.0 has an improved question set, a new tool for datasheet completion, and improved instructions and completion guidance, helping users to complete the datasheet more consistently and comparably. We make all HEDS 3.0 resources available online.<sup>1</sup>

## 1 Introduction

The Human Evaluation Datasheet (HEDS), first introduced in 2021 (Shimorina and Belz, 2021), is conceived as a template for recording and reporting the details of human evaluation experiments in a standardised and comparable way with NLP-wide scope. It has been extensively used in practice, in particular in the context of the Re-proGen/ReproNLP shared task series (Belz et al., 2021, 2022; Belz and Thomson, 2023, 2024; Belz et al., 2025c), where organisers and participants have been completing HEDS sheets for original studies and reproduction studies, respectively.<sup>2</sup>

This in turn has provided new insights into what information HEDS needs to capture, what functionality is needed in an interactive tool for its completion, and what guidance needs to be provided to users to enable them to complete HEDS sheets quickly and consistently. We have channelled these insights into a new version update of HEDS, numbered 3.0 which has (i) major updates to questions and answers, (ii) new resources provided as part of the HEDS 3.0 package, and (iii) improved detail and clarity in the user guidance.

Re *i*, we have added two new questions, and replaced seven questions with two or more specific

ones each. Re *ii*, we have replaced the original Google form with the tailored interactive HEDS 3.0 tool which supports browsing, revision, pre-filling of some questions, and exporting to LaTeX and JSON. Re *iii*, we have revised, extended and improved the clarity of completion instructions and incorporated them into the HEDS 3.0 tool.

The paper is structured as follows. We summarise contributions to previous versions of HEDS on which HEDS 3.0 is based (Section 2). We present an overview of HEDS 3.0 in terms of the components that make up the HEDS 3.0 package in Section 3.1, followed by a description of question types and presentational conventions (Section 3.2). Section 3.3 presents the parts of the instructions from the HEDS 3.0 tool that relate to the content of the form (omitting those relating to technical aspects of the tool only). A summary of differences between questions in HEDS 3.0 vs. HEDS 2.0 can be found in Section 3.4.

Section 4 gives an overview of the HEDS 3.0 tool, and Section 5 describes envisaged uses of HEDS. In Section 6 we provide additional explanations for some aspects of HEDS 3.0 that we know from experience users may find more difficult. We end with some discussion and conclusions in Section 7. The complete HEDS 3.0 sheet is included in the appendix, as a printout of questions and possible answers automatically generated from the version of the sheet used in the HEDS 3.0 tool (Appendix A).

## 2 Credits

HEDS 1.0 (2021) and HEDS 2.0 (2022) were created by Shimorina and Belz who in turn acknowledge the following sources: Questions 2.1–2.5 relating to evaluated system(s), and 4.3.1–4.3.8 relating to response elicitation, ultimately derive from 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

<sup>1</sup><https://github.com/DCU-NLG/HEDS-3.0>

<sup>2</sup><https://repronlp.github.io>.

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) and van der Lee et al. (2021), and by Gehrmann et al. (2021)’s 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).

The questions newly added in HEDS 3.0 (see Section 3.4) were created by the authors of this paper to address documentation needs that arose primarily in the context of the ReproHum Project and related ReproNLP shared task series (Belz and Thomson, 2023, 2024).<sup>3</sup> For example, whereas Q3.2.2 previously asked a single broad question about the type of evaluators used; there are now separate questions for domain expertise (Q3.2.2.1), payment (Q3.2.2.2), and whether the participants were authors (Q3.2.2.4), or previously known to the authors (Q3.2.2.3) (for full listing see Section 3.4).

### 3 HEDS 3.0 Overview

#### 3.1 Package components

The HEDS 3.0 package consists of the following three resources, all accessible via <https://github.com/DCU-NLG/HEDS-3.0>:

1. The HEDS 3.0 tool comprising the interactive form and instructions for completion: available for online completion at <https://nlp-heds.github.io>;
2. Description and completion guidance: this document and on GitHub;
3. Scripts for exporting completed HEDS 3.0 forms to alternative formats, including Latex:<https://github.com/DCU-NLG/HEDS-3.0>.

#### 3.2 Structure, question types and presentation

HEDS is divided into five sections as follows:

1. Main Reference and Supplementary Resources (Questions 1.1.1–1.3.2.3);
2. Evaluated System(s) (Questions 2.1–2.5);
3. Sample of System Outputs, Evaluators and Experimental Design (Questions 3.1.1–3.3.8);

4. Definition and Operationalisation of Quality Criteria (Questions 4.1.1–4.3.12.2);
5. Ethics (Questions 5.1–5.4).

In Appendix A we present the HEDS 3.0 form in its entirety, in a similar look/feel to the online version that users complete (in fact, the whole section is generated automatically from the form).

Questions come in the following types and presentation formats:

1. Multiple-choice list, select one: radio buttons. For example, Question 4.2.2 asks “Are outputs assessed in absolute or relative terms?”, with response options of “absolute” or “relative”.
2. Multiple-choice list, select all that apply: check boxes. For example, Question 2.5 asks “What are the language(s) of the outputs produced by the system?”, with response options taken from the list of standardised full language names as per ISO 639-1 (2019). The options “N/A” and “Other” are also available, with a text box appearing if they are selected that allows for responses to be explained or described.
3. Short text box, enter one type of information (a URL, a value range, etc.). For example, Question 4.3.1.1 asks “What do you call the quality criterion in explanations/interfaces to evaluators?”. As a name, this does not require more than a single line of text.
4. Longer text box: enter (a) more comprehensive information, and/or (b) information that depends on given factors. For example, Question 4.3.2 asks “What definition do you give for the quality criterion in explanations/interfaces to evaluators?”. Depending on the quality criterion, this may require a longer definition.

#### 3.3 Instructions

The following text is presented at the start of completing the online HEDS 3.0 form, to support users in answering the questions in it. The verbatim text shown below was generated automatically from the form (except for the insertion of subsection headers).

*Text of instructions generated by HEDS 3.0 tool:*

This is the Human Evaluation Datasheet (HEDS) form which is designed to record full details of human evaluation experiments in Natural Language

<sup>3</sup><https://reprohum.github.io>

Processing (NLP), addressing a history of details often going unreported in the field (in extreme cases, no details at all are reported). Reporting such details is crucial for gauging the reliability of results, determining comparability with other experiments, and for assessing reproducibility (Belz et al., 2023a,c; Thomson et al., 2024; Thomson and Belz, 2024). Having a standard set of questions to answer (as provided by HEDS) means not having to worry about what information to include or in what detail, as well as the information being in a format directly comparable to information reported for other human evaluation experiments. To maximise standardisation, questions are in multiple-choice format where possible.

The HEDS form is divided into five main sections, containing questions that record information about resources, evaluated system(s), test set sampling, quality criteria assessed, and ethics, respectively. Within each of the main sections there can be multiple subsections which can be expanded or collapsed.

Each HEDS question comes with instructions and notes to help with answering it, except where the task is exceedingly simple (e.g. when a contact email address is asked for).

HEDS Section 4 needs to be completed for each quality criterion that is evaluated in the experiment. Instructions on how to do this are shown at the start of HEDS Section 4.

The form is not submitted to any server when it is completed, and instead needs to be downloaded to a local file. A tool is available in the GitHub repository for converting the file to latex format (which we used to generate the next section).

We recognise that completing a form of this length and level of detail constitutes an overhead in terms of time and effort, especially the first time a HEDS form is completed when the learning curve is steepest. However, this overhead does go down substantially with each use of HEDS, and, we believe, is far outweighed by the benefits: increased scientific rigour, reliability and repeatability.

### 3.4 Changes to questions compared to HEDS 2.0

We have introduced two new questions (4.3.12.1 and 4.3.12.2), and have in seven cases replaced what was a single question in HEDS 2.0 with two or more in 3.0. For example, there was one question on inter-annotator agreement in 2.0 (4.3.11), whereas now there are two (4.3.11.1 and

4.3.11.2). All questions with numbering of depth 4 (e.g. 4.3.11.1), and two of depth 3, are the result of such a replacement. In some cases, the motivation was to accommodate a new question without changing other question numbers. In other cases, it was to split an existing question into two for increased clarity and consistency. The complete list of question number mappings from version 2.0 to version 3.0 is as follows:

1.1	→	1.1.1, 1.1.2
1.3	→	1.3.1.1, 1.3.1.2, 1.3.1.3, 1.3.2.1, 1.3.2.2, 1.3.2.3
3.1.3	→	3.1.3.1, 3.1.3.2, 3.1.3.3
3.2.2	→	3.2.2.1, 3.2.2.2, 3.2.2.3, 3.2.2.4
3.3.3	→	3.3.3.1, 3.3.3.2
3.3.4	→	3.3.4.1, 3.3.4.2
4.3.11	→	4.3.11.1, 4.3.11.2
–	+	4.3.12.1, 4.3.12.2

For each of the eight lines above, we explain the change and the motivation for it below:

Q1.1: Previously, Question 1.1 captured the “link to paper reporting the evaluation experiment,” *and* asked the user to “state which experiment you’re completing this sheet for.” We replace it with two questions Q1.1.1 and Q1.1.2 in order to separate the two details.

Q1.3: Question 1.3 captured “name, affiliation and email address of person completing this sheet, and of contact author if different.” in a single text box. We replace it with separate questions for the name, affiliation, and email address of the person completing the sheet (Q1.3.1.1, Q1.3.1.2, and Q1.3.1.3 respectively) as well as for the the contact author (Q1.3.2.1, Q1.3.2.2, and Q1.3.2.3).

Q3.1.3: Previously, Question 3.1.3 captured “the results of a statistical power calculation on the output sample,” *and* asked the user to “provide numerical results and a link to the script used.” We replace it with three separate questions, Q3.1.3.1 (recording the method used), Q3.1.3.2 (recording the statistical power value) and Q3.1.3.3 (recording a link to the code).

Q3.2.2: Question 3.2.2 captured what “kind of evaluators are in this experiment.” However, the user was also asked to “In all cases, provide details in the text box under *Other*.” To separate these issues, we replace Q3.2.2 with

Figure 1: Screenshot of the web-based HEDS 3.0 tool.

Q3.2.2.1 (whether participants are domain experts), Q3.2.2.2 (whether participants received any form of payment), Q3.2.2.3 (whether participants were previously known to authors), and Q3.2.2.4 (whether any authors were also participants). This removed the issue of having one question ask for multiple things and also prompts the user to consider specific important characteristics of the evaluators.

Q3.3.3: Question 3.3.3 captured the “quality assurance methods [that] are used”. However, the user was also asked to “In all cases, provide details in the text box under *Other*.”. We replace this with Q3.3.3.1 (recording the types of quality assurance methods are used) and Q3.3.3.2, which records the methods that are used for each of the types of quality assurance methods that were selected in Question 3.3.3.1. Q3.3.3.1 is a multiple choice list, allowing for the user to select from a list of clearly defined methods (or enter “Other” and specify). This can then be elaborated in Q3.3.3.2.

Q3.3.4: Question 3.3.4 captured what “evalua-

tors see when carrying out evaluations.” However, it asked the user to “link to screenshot(s)” *and/or* “describe the evaluation interface(s).” We split this into two questions, with Q3.3.4.1 capturing the link and Q3.3.4.2 a description.

Q4.3.11: Question 4.2.11 asked “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?” We first separate inter from intra-annotator agreement (4.3.11.\* and 4.3.12.\* respectively). For each we now capture the method (4.3.11.1, 4.3.12.1) and the score (4.3.11.2, 4.3.12.2).

–: See previous bullet re the introduction of Questions 4.3.12.1 and 4.3.12.2.

All questions now ask for a single piece of information (some having the option of an elaboration for certain response options). This both clearly separates the recorded information and reduces the chance of the user omitting information. In all other cases, questions are in essence the same (apart from rewording), and have the same number, in both versions, apart from the minor respects noted below.

*Question wording:* Most questions have undergone some degree of rewording in order to make them (a) clearer and easier to answer, and (b) more consistent in wording and style.

*Answer types:* In a small number of cases we have replaced a text box answer with a list of options, to achieve greater comparability in answers between users.

The overall motivation for all changes was to make it easier for users to complete the datasheet consistently and comparably (to other users).

## 4 The HEDS 3.0 Tool

A web-based version of HEDS 3.0 has been implemented in HTML and Javascript. It can be accessed for online completion,<sup>4</sup> or alternatively, users can download the code<sup>5</sup> and run it on their own computer.

Figure 1 shows a screenshot of the HEDS 3.0 tool homepage. The sidebar to the left contains:

- A button to download a JSON file containing the form contents (which are otherwise stored in the web browser cache). It is this file which can be used to generate the LaTeX format output using the python script that we provide.<sup>6</sup>
- A file upload section to load form contents for such a JSON file.
- A section showing a count of errors such as fields which are blank, or errors where invalid multiple choice combinations have been selected.

The main body of the form has seven expandable headers. First there is the *Introduction*, which explains what HEDS is and how to use the form. Then are five numbered sections that correspond to the numbered HEDS sections as shown in Section 3.2 and as can also be seen in Appendix A. When expanded, these sections contain further expandable headers and ultimately, questions. For example, in Figure 2, the Section 1 header and then the 1.3 and 1.3.1 headings have been expanded, revealing the three Q1.3.1.\* questions which record the details of the person who is completing the sheet.

<sup>4</sup><https://nlp-heds.github.io/>

<sup>5</sup><https://github.com/DCU-NLG/HEDS-3.0>

<sup>6</sup>(Appendix A is simply a blank form generated using said script.

Figure 2: Screenshot of web-based HEDS 3.0 tool with Sections 1, 1.3, and 1.3.1 expanded to show Questions 1.1.3.1–1.1.3.3. The warning messages disappear once the information has been entered.

Section 4 of the HEDS form is completed for each quality criterion that is being evaluated. Figure 3 shows how the web tool handles this; by creating a new tab per quality criterion.

Finally, there is the *All Form Errors* section (bottom left of Figure 1) which when expanded will show the numbers of all questions that have errors.

## 5 Envisaged uses

We envisage the main uses of HEDS to be as follows.

### 5.1 Preregistration

Ideally, HEDS should be completed before a human evaluation experiment is run, at the point when the design is final, as part of a formal preregistration process. The preregistration documents submitted can then include the completed HEDS form.

After that point, the experimental design, and therefore the HEDS sheet, should no longer be changed. Once the experiment has been run, the information in the sheet can be updated if necessary, e.g. if the final number of evaluators had to change due to unforeseen circumstances.



### Many Criteria : Quality Criterion - Definition and Operationalisation

In this section you can create named subsections for each criterion that is being evaluated. The form is then duplicated for each criterion. To create a criterion type its name in the field and press the *New* button, it will then appear on tab that will allow you to toggle the active criterion. To delete the current criterion press the *Delete current* button.

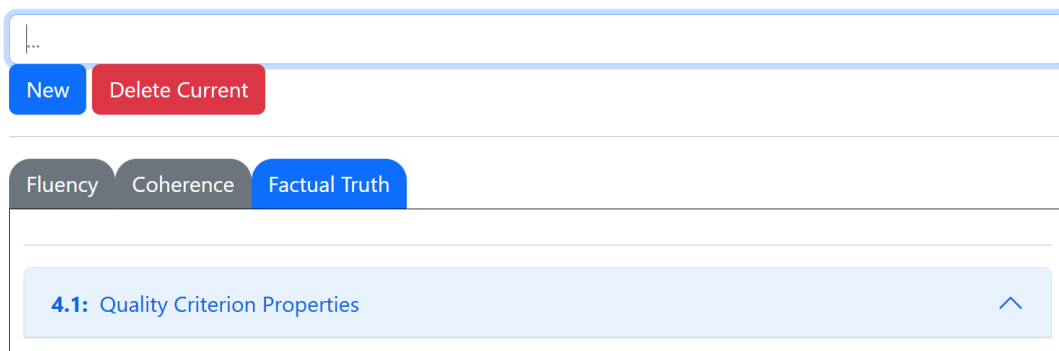


Figure 3: Screenshot showing how multiple quality criteria can be added in Section 4 of web-based HEDS 3.0.

## 5.2 Reporting

Another use is for the purpose of reporting the details of a completed experiment. For this, the completed HEDS sheet can be automatically converted to Latex, ready for inclusion in the supplementary material or appendix of the paper reporting the experiment.

The advantage in reporting this information in standardised form is ensuring that complete and directly comparable information is recorded for human evaluation studies, in turn helping reproducibility.

## 5.3 Reproduction studies

A third use is in carrying out reproducibility studies, where the properties of the original study are captured in a HEDS sheet and reproduction studies are implemented so as to have the same properties.

This has been done extensively in the ReproGen and ReproNLP shared tasks (Belz et al., 2022; Belz and Thomson, 2024). Here, the HEDS sheets were used to ensure that original work and reproduction experiment had the same properties, hence can be expected to produce similar results.

## 6 Additional Explanations

### *Meaning of ‘experiment’*

In the context of HEDS, an experiment consists of a set of assessments for one or more evaluation methods each assessing one quality criterion, that are collected at the same time, with the same experimental design. This means that for a given experiment, all HEDS questions except for those

in HEDS Section 4 (about quality criteria) need to be answered only once.

*Question 4.3.1.2: What standardised quality criterion name does the name entered for 4.3.1.1 correspond to?*

As discussed in detail elsewhere (Howcroft et al., 2020; Belz et al., 2025a), just because two evaluation experiments use the same quality criterion name does not mean that they assess the same aspect of quality. The only way we can be sure that the same aspect of quality is being assessed is if we map the two quality criterion names to a single standard set of quality criteria via the same systematic mapping process.

The QCET taxonomy of quality criteria (Belz et al., 2025a) was designed to provide both a standard set of quality criteria names and definitions, and the mapping process. It does this via the taxonomic structure which is intended to be followed top down on the way to identifying the node that best matches the quality criterion name that is to be standardised.

By using the standardised quality criteria from QCET, one can also identify for each quality criterion, the correct type of quality assessed (Question 4.1.1), aspect of system outputs assessed (Question 4.1.2), and the frame of reference (Question 4.1.3). These pieces of information are fixed for each QCET quality criterion and can be seen when viewing a quality criterion node in the taxonomy.

## 7 Discussion and Conclusion

It is the norm (Belz et al., 2023b) in NLP to publish very little detail about human evaluations, with complete sharing of details practically unheard of (Thomson and Belz, 2024). This is true even in cases where major conclusions in a paper depend on the results. For example, it is quite common to mention just the number of evaluators used, and the quality criteria assessed, before presenting tables of mean ratings. Clearly, in this situation it's not possible to assess whether the evaluation is sound, the methods of analysis applied are appropriate, or conclusions supported.

Moreover, without publishing details of human evaluations, it can't be established whether two evaluations assess the same thing, thus whether they agree with each other or not in their assessment of different types of systems. Without that, our ability to build on results, to progress collectively as a field of science, is greatly reduced (Jones, 1981).

Diligent reporting always represents an overhead in terms of effort, one that in the fast moving field of NLP it is tempting to avoid. However, the more impactful NLP (and AI more generally) becomes, the more important it is that it adopts scientific practices, and reporting full details of evaluations is an important part of that.

With HEDS, our aim is to contribute to this change, reducing the load on researchers somewhat by making it possible to report full details about a human evaluation by completing an interactive form, then exporting a fully formatted PDF that can simply be attached as an appendix or supplementary material of the paper reporting the work. It can also be exported to JSON format for use in automatic comparison between multiple evaluations for use in e.g. comparability and reproducibility assessments.

## Acknowledgments

Thomson's contribution was funded by the ADAPT SFI Centre for Digital Media Technology. Our work has also benefited more generally from being carried out within the research environment of the ADAPT SFI Centre, funded by Science Foundation Ireland through the SFI Research Centres Programme and co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106.

## Appendix

### A HEDS Form in its Entirety

#### HEDS Section 1: Main Reference and Supplementary Resources

##### 1.1 Main reference

**Question 1.1.1: Where can the main reference for the evaluation experiment be found?**

*Multiple-choice options (select one):*

- ☐ *The main paper reporting the experiment is here (enter URL).*
- ☐ *An unpublished report describing the experiment can be found here (enter URL).*
- ☐ *No report describing the experiment is available and this sheet will be uploaded for preregistration here (enter URL).*
- ☐ *No report describing the experiment is available and no pre-registration is not planned.*

**Question 1.1.2: Which experiment is this form being completed for?**

*What to enter in the text box:* Referring to the main reference entered for Question 1.1.1, identify the experiment that you're completing this form for (see instructions section at the start for explanation of term 'experiment'), in particular to differentiate this experiment from any others that you are carrying out as part of the same overall work: (a) if a link for a published paper was entered under Question 1.1.1, give here the section(s) and/or table(s) that best identify the experiment, plus a brief description for clarity; (b) if 'preregistration' or 'unpublished' was selected, enter a brief description of the experiment, mentioning quality criteria, dataset and systems.

##### 1.2 Supplementary resources

**Question 1.2: Where can the resources that were used in the evaluation experiment be found?**

*Multiple-choice options (select one):*

- *The resources used in the experiment can be found here (enter URL(s)).*
- *No resources shared.*

### 1.3 Contact Details

#### 1.3.1 Details of the person completing this sheet.

**Question 1.3.1.1: Name of the person completing this sheet.**

**Question 1.3.1.2: Affiliation of the person completing this sheet.**

**Question 1.3.1.3: Email address of the person completing this sheet.**

#### 1.3.2 Details of the contact author

**Question 1.3.2.1: Name of the contact author.**

**Question 1.3.2.2: Affiliation of the contact author.**

**Question 1.3.2.3: Email address of the contact author.**

### HEDS Section 2: Evaluated System(s)

*Notes:* Questions 2.1–2.5 in this section record information about the system(s) that are evaluated in the experiment this sheet is being completed for. The input, output and task questions are closely interrelated: the answer to one partially determines the answer to the others, as indicated for some combinations of answers under Question 2.3.

**Question 2.1: What type of input do the evaluated system(s) take?**

*Notes:* The term ‘input’ here refers to the text, representations and/or data structures that all of

the evaluated systems take as input (including prompts). 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. E.g. the input to Referring Expression Generation (REG), end-to-end text generation, etc. NB: excludes linguistic representations.
- ☐ **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:** Unit(s) of text shorter than a sentence, e.g. Referring Expressions (REs), verb phrase, text fragment of any length; includes titles/headlines.
- ☐ **Text: sentence:** Single sentence(s).
- ☐ **Text: multiple sentences:** Sequence(s) of multiple sentences, without any document structure.
- ☐ **Text: document:** Text(s) with document structure, such as a title, paragraph breaks or sections, e.g. a set of news reports for summarisation.
- ☐ **Text: dialogue:** Dialogue(s) of any length, excluding a single turn which would come under one of the other text types.
- ☐ **Text: other (please describe):** Input is text but doesn’t match any of the above text categories.
- ☐ **Speech:** Recording(s) of speech.
- ☐ **Visual:** Image(s) or video(s).
- ☐ **Multi-modal:** Select this option if input is *always* a combination of multiple modalities. Also select other options in this list to different elements of the multi-modal input.
- ☐ **Control feature:** Feature(s) or parameter(s) specifically present to control a property of the output text, e.g. positive stance, formality, author style.



- ☐ **No input (please explain):** If there are no system inputs, select this option and explain why.
- ☐ **Other (please describe):** If input is none of the above, select this option and describe it.

**Question 2.2: What type of output do the evaluated system(s) generate?**

*Notes:* The term ‘output’ here refers to the text, representations and/or data structures that all of the evaluated systems produce as output. 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.

*Check-box options (select all that apply):*

- ☐ **Raw/structured data:** Numerical, symbolic, and other data, possibly structured into trees, graphs, graphical models, etc. E.g. the input to Referring Expression Generation (REG), end-to-end text generation, etc. NB: excludes linguistic representations.
- ☐ **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:** Unit(s) of text shorter than a sentence, e.g. Referring Expressions (REs), verb phrase, text fragment of any length; includes titles/headlines.
- ☐ **Text: sentence:** Single sentence(s).
- ☐ **Text: multiple sentences:** Sequence(s) of multiple sentences, without any document structure.
- ☐ **Text: document:** Text(s) with document structure, such as a title, paragraph breaks or sections, e.g. a set of news reports for summarisation.
- ☐ **Text: dialogue:** Dialogue(s) of any length, excluding a single turn which would come under one of the other text types.
- ☐ **Text: other (please describe):** Input is text but doesn’t match any of the above text categories.

- ☐ **Speech:** Recording(s) of speech.
- ☐ **Visual:** Image(s) or video(s).
- ☐ **Multi-modal:** Select this option if input is *always* a combination of multiple modalities. Also select other options in this list to different elements of the multi-modal input.
- ☐ **No input (please explain):** If there are no system inputs, select this option and explain why.
- ☐ **Other (please describe):** If input is none of the above, select this option and describe it.

**Question 2.3: What is the task that the evaluated system(s) perform in mapping the inputs in Question 2.1 to the outputs in Question 2.2?**

*Notes:* This question is about the task(s) 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 features* 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 system task corresponds to more than one of tasks above, but the system doesn't implement them as separate tasks.
- ☐ **Image/video description:** Input includes *visual*, and the output describes it in some way.
- ☐ **Post-editing/correction:** The system edits and/or corrects the input text (can itself be the textual output from another system) to yield an improved version of the text.
- ☐ **Other (please describe):** If task is none of the above, Select this option and describe it.

**Question 2.4: What are the language(s) of the inputs accepted by the system(s)?**

*Notes:* Select any language(s) that apply from this list of standardised full language names as per [ISO 639-1 \(2019\)](#). If language is not (part of) the input, select 'N/A'.

*Check-box options (select all that apply):*

- ☐ **N/A (please explain):** No language in the input.
- ☐ **Abkhazian:** Also known as Abkhaz.
- ☐ **Afar.**
- ☐ **Afrikaans.**
- ☐ ...
- ☐ **Zhuang, Chuang.**
- ☐ **Zulu.**
- ☐ **Other (please describe):** A language that is not on the above list.

**Question 2.5: What are the language(s) of the outputs produced by the system?**

*Notes:* Select any language(s) that apply from this list of standardised full language names as per [ISO 639-1 \(2019\)](#). If language is not (part of) the output, select 'N/A'.

*Check-box options (select all that apply):*

- ☐ *N/A (please explain):* No language is generated.
- ☐ *Abkhazian:* Also known as Abkhaz.
- ☐ *Afar.*
- ☐ *Afrikaans.*
- ☐ ...
- ☐ *Zhuang, Chuang.*
- ☐ *Zulu.*
- ☐ *Other (please describe):* A language that is not on the above list.

**HEDS Section 3: Sample of system outputs, evaluators, experimental design**

**3.1 Sample of system outputs (test set)**

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?**

*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. For most experiments this should be a single integer. If the number of outputs varies please explain how and why.

**Question 3.1.2: How are system outputs (or other evaluation items) selected for inclusion?**

*Multiple-choice options (select one):*

- ☐ *By simple automatic random selection:* Outputs are selected from a larger set by a script using a pseudo-random number generator, without stratification, every-*n*th selection, etc.

- ☐ *By an automatic random process but using stratified sampling over given properties:* Selection is by a random script as above, but with added constraints ensuring that the sample is representative of the set of outputs it is selected from, in terms of given properties, such as sentence length, positive/negative stance, etc.
- ☐ *By non-random automatic selection:* Output sample is selected by a non-randomised automatic process, e.g. selecting every *n*th item.
- ☐ *By manual, arbitrary selection:* Output sample was selected by hand, or automatically from a manually compiled list, without specific selection criteria.
- ☐ *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 describe):* If selection method is none of the above, select this option and describe it.

**3.1.3 Statistical power of the sample**

*Notes:* All evaluation experiments should perform a power analysis to determine an appropriate sample size. If none was performed, enter 'N/A' in Questions 3.1.3.1–3.1.3.3

**Question 3.1.3.1: What method of statistical power analysis was used to determine the appropriate sample size?**

*What to enter in the text box:* The name of the method used, and a URL linking to a reference for the method.

**Question 3.1.3.2: What is the statistical power of the sample?**

*What to enter in the text box:* The numerical results of the statistical power calculation on the output sample obtained with the method in Question 3.1.3.1.

**Question 3.1.3.3: Where can other researchers find details of any code used in the power analysis performed?**

*What to enter in the text box:* A URL linking to any code used in the calculation in Question 3.1.3.2.

### 3.2 Evaluators

**Question 3.2.1: How many evaluators are there in this experiment?**

*What to enter in the text box:* A single integer representing the total number of evaluators whose assessments contribute to results in the experiment. Don't count evaluators who performed some evaluations but who were subsequently excluded.

#### 3.2.2 Evaluator Type

**Question 3.2.2.1: Are the evaluators in this experiment domain experts?**

*Multiple-choice options (select one):*

- ☐ **Yes:** Participants are considered domain experts, e.g. meteorologists evaluating a weather forecast generator, or nurses evaluating an ICU report generator.
- ☐ **No:** Participants are not domain experts.
- ☐ **N/A (please explain).**

**Question 3.2.2.2: Did participants receive any form of payment?**

*Multiple-choice options (select one):*

- ☐ ***Paid (monetary compensation):*** Participants were given some form of monetary compensation for their participation.
- ☐ ***Paid (non-monetary compensation such as course credits):*** Participants were given some form of non-monetary compensation for their participation, e.g. vouchers, course credits, or reimbursement for travel unless based on receipts.
- ☐ ***Not paid:*** Participants were not given compensation of any kind (except for receipt-based reimbursement of expenses).
- ☐ ***N/A (please explain).***

**Question 3.2.2.3: Were any of the participants previously known to the authors?**

*Multiple-choice options (select one):*

- ☐ **Yes:** One or more of the researchers running the experiment knew some or all of the participants before recruiting them for the experiment.
- ☐ **No:** None of the researchers running the experiment knew any of the participants before recruiting them for the experiment.
- ☐ **N/A (please explain).**

**Question 3.2.2.4: Were any of the researchers running the experiment among the participants?**

*Multiple-choice options (select one):*

- ☐ **Yes:** Evaluators include one or more of the researchers running the experiment.
- ☐ **No:** Evaluators do not include any of the researchers running the experiment.
- ☐ **N/A (please explain).**

**Question 3.2.3: How are evaluators recruited?**

*What to enter in the text box:* 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?

**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:* Describe any training evaluators were given to prepare them for the evaluation task, including any practice evaluations they did. This includes introductory explanations, e.g. on the start page of an online evaluation tool.

**Question 3.2.5: What other characteristics do the evaluators have?**

*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, e.g. because of information collected during the evaluation. This might include geographic location, 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.

### 3.3 Experimental Design

**Question 3.3.1: Has the experimental design been preregistered?**

*Notes:* If the answer is yes, also give a link to the registration page for the experiment.

*Multiple-choice options (select one):*

- ☐ *Yes (please provide link).*
- ☐ *No.*

**Question 3.3.2: By what medium are responses collected?**

*What to enter in the text box:* Describe the platform or other medium used to collect responses, e.g. paper forms, Google forms, SurveyMonkey, Mechanical Turk, CrowdFlower, audio/video recording, etc.

#### 3.3.3 Quality assurance

*Notes:* Question 3.3.3.1 records information about the *type(s)* of quality assurance employed, and Question 3.3.3.2 records the details of the corresponding quality assurance methods.

**Question 3.3.3.1: What types of quality assurance methods are used to ensure that evaluators are sufficiently qualified and/or their responses are of sufficient quality?**

If any quality assurance methods other than those listed were used, select ‘other’, and describe why below. If no methods were used, select *none of the above*.

*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 and/or after evaluation:*** Evaluations are checked for quality by automatic scripts during or after evaluations, e.g. evaluators are given known bad/good outputs to check that scores are appropriate.
- ☐ ***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.
- ☐ ***Other (please describe):*** Briefly mention any other quality-assurance methods that were used. Details of the method should be entered under 3.3.3.2.
- ☐ ***None of the above (no quality assurance methods used).***

**Question 3.3.3.2: What methods are used for each of the types of quality assurance methods that were selected in Question 3.3.3.1?**

*What to enter in the text box:* Give details of the methods used for each of quality assurance types from the last question. E.g. if quality checks were used, give details of the check. If no quality assurance methods were used, enter ‘N/A’.



### 3.3.4 Form/Interface

**Question 3.3.4.1: Where can the form/interface that was shown to participants be viewed?**

*What to enter in the text box:* Enter a URL linking to a screenshot or copy of the form if possible. If there are many files, please create a signpost page (e.g. on [GitHub](#)) that contains links to all applicable files. If there is a separate introductory interface/page, include it under Question 3.2.4.

**Question 3.3.4.2: What types of information are evaluators shown when carrying out evaluations?**

*What to enter in the text box:* Describe the types of information (the evaluation item, a rating instrument, instructions, definitions, etc.) evaluators can see while carrying out each assessment. In particular, explain any variation that cannot be seen from the information linked to in Question 3.3.4.1.

**Question 3.3.5: How free are evaluators regarding when and how quickly to carry out evaluations?**

*Check-box options (select all that apply):*

- ☐ **Evaluators must carry out the evaluation at a specific time/date.**
- ☐ **Evaluators must complete each individual assessment within a set amount of time.**
- ☐ **Evaluators must complete the whole evaluation within a set amount of time.**
- ☐ **Evaluators must complete the whole evaluation in one sitting:** Partial progress cannot be saved and the evaluation cannot be returned to on a later occasion.
- ☐ **None of the above (please describe):** Select this option if none of the above are the case in the experiment, then describe any other constraints imposed on when and/or how quickly evaluations must be carried out.

**Question 3.3.6: Are evaluators told they can ask questions about the evaluation and/or provide feedback?**

*Check-box options (select all that apply):*

- ☐ **Evaluators can ask questions during 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 while carrying out their assessments.
- ☐ **Evaluators provide feedback after the evaluation:** Evaluators are explicitly asked to provide feedback and/or comments about the evaluation after completing it, either verbally or in written form, e.g. via an exit questionnaire or a comment box.
- ☐ **Other (please describe):** Use this space to describe any other ways you provide for evaluators to ask questions or provide feedback.
- ☐ **None of the above:** Select this option if evaluators are not able to ask questions or provide feedback.

**Question 3.3.7: What are the conditions in which evaluators carry out the evaluations?**

*Multiple-choice options (select one):*

- ☐ **Evaluators carry out assessments at a place of their own choosing:** Evaluators are given access to the evaluation medium specified in Question 3.3.2, and subsequently choose where to carry out their evaluations.
- ☐ **Evaluators carry out assessments in a lab, and conditions are controlled to be the same for each evaluator.**
- ☐ **Evaluators carry out assessments in a lab, and conditions are not controlled to be the same for different evaluators.**
- ☐ **Evaluators carry out assessments in a real-life situation, and conditions are controlled to be 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.

- *Evaluators carry out assessments in a real-life situation, and conditions are not controlled to be the same for different evaluators.*
- *Evaluators carry out assessments outside of the lab, in a situation designed to resemble a real-life situation, and conditions are controlled to be 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.
- *Evaluators carry out assessments outside of the lab, in a situation designed to resemble a real-life situation, and conditions are not controlled to be the same for different evaluators.*
- *Other (please describe):* Use this space to provide additional, or alternative, information about the conditions in which evaluators carry out assessments, not covered by the options above.

**Question 3.3.8: In what ways do conditions in which evaluators carry out the evaluations vary for different evaluators?**

*What to enter in the text box:* For those conditions that are not controlled to be the same, describe the variation that can occur. For conditions that are controlled to be the same, enter 'N/A'.

## HEDS Section 4: Definition and Operationalisation of Quality Criteria

*Notes:* Questions in this section record information about each quality criterion (Fluency, Grammaticality, etc.) assessed in the human evaluation experiment that this sheet is being completed for.

If multiple quality criteria are evaluated, the form creates subsections for each criterion headed

by the criterion name for each one. These are implemented as overlaid windows with tabs for navigating between them.

### 4.1 Quality Criterion Properties

*Notes:* Questions 4.1.1–4.1.3 capture aspects of quality assessed by a given quality criterion in terms of three orthogonal properties: (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, 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:** Select 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:** Select 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 non-sensical 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:** Select this option if the criterion assesses outputs as a whole, not just form or just content. E.g. Coherence, Usefulness and Task Completion 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:** Select 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:** Select 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.

## 4.2 Evaluation mode properties

*Notes:* Questions 4.2.1–4.2.3 record properties that are orthogonal to quality criterion properties (preceding section), i.e. any given quality criterion can in principle be combined with any of the modes (although some combinations are much 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:** Select 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 should yield the same score/result.
- **Subjective:** Select 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:** Select 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:** Select this option if quality of outputs is assessed *without* considering their effect on something external to the system such as the performance of an embedding system or of a user at a task.
- **Extrinsic:** Select 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.

### 4.3 Response elicitation

*Notes:* 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. This includes what is presented to evaluators, how they select a response, and via what type of tool, etc.

#### 4.3.1 Quality criterion name

**Question 4.3.1.1: What do you call the quality criterion in explanations/interfaces to evaluators?**

*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 ‘no name given’.

**Question 4.3.1.2: What standardised quality criterion name does the name entered for 4.3.1.1 correspond to?**

*What to enter in the text box:* Map the quality criterion name used in the evaluation experiment to its equivalent in a standardised set of quality criterion names and definitions such as QCET (Belz et al., 2024, 2025b), and enter the standardised name and reference to the paper here. In performing this mapping, the information given in Questions 4.3.7 (question/prompt), 3.3.4.1–3.3.4.2 (interface/information shown to evaluators), 4.3.2 (QC definition), 3.2.4 (training/practice), and 4.3.1.1 (verbatim QC name) should be taken into account, in this order of precedence.

**Question 4.3.2: What definition do you give for the quality criterion in explanations/interfaces to evaluators?**

*What to enter in the text box:* Copy and paste 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 ‘no definition given’.

**Question 4.3.3: What is the size of the scale or other rating instrument?**

*What to enter in the text box:* An integer representing the number of different possible response values obtained with the scale or rating instrument. Enter ‘continuous’ if the number of response values is not finite. Enter ‘N/A’ if there is no scale or rating instrument. E.g. for a 5-point rating scale, enter ‘5’; for a slider that can return 100 different values (even if it looks continuous), enter ‘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: What are the possible values of the scale or other rating instrument?**

*What to enter in the text box:* List, or give the range of, the possible response values returned by 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 collected via a slider, the list entered might be ‘[-50,+50]’. If no rating instrument is used, enter ‘N/A’.

**Question 4.3.5: How is the scale or other rating instrument presented to evaluators?**

*Multiple-choice options (select one):*

- ☐ **Multiple-choice options:** Select this option if evaluators select exactly one of multiple options.
- ☐ **Check-boxes:** Select this option if evaluators select any number of options from multiple given options.
- ☐ **Slider:** Select 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):** Select this option if there is no rating instrument.
- ☐ **Other (please describe):** Select 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, what is the task the evaluators perform?**

*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, use this space to describe the task evaluators perform, and what information is recorded. Tasks that don't use rating instruments include ranking multiple outputs, finding information, playing a game, etc.). If there is a rating instrument, enter 'N/A'.

**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: What form of response elicitation is used in collecting assessments from evaluators?**

The terms and explanations in this section have been adapted from [Howcroft et al. \(2020\)](#).

*Multiple-choice options (select one):*

- **(Dis)agreement with quality statement:** Participants indicate the degree to which they agree with a given quality statement 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 indicate level of quality on 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:** Select 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:** Select 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:** Select 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:** Select 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:** Select 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 describe):** 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?**

*What to enter in the text box:* Normally a set of separate assessments is collected from evaluators and then 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 summarising, per-system results. If no such method was used, enter 'results were not processed or aggregated before being reported'.

**Question 4.3.10: What method(s) are used for determining effect size and significance of findings for this quality criterion?**

*What to enter in the text box:* The 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, enter 'None'.

#### 4.3.11 Inter-annotator agreement

**Question 4.3.11.1: How was the inter-annotator agreement between evaluators measured for this quality criterion?**

*What to enter in the text box:* The method(s) used for measuring inter-annotator agreement. If inter-annotator agreement was not measured, enter 'InterAA not assessed'.

**Question 4.3.11.2: What was the inter-annotator agreement score?**

*What to enter in the text box:* The inter-annotator agreement score(s) obtained with the method(s) in Question 4.3.11.1. Enter 'InterAA not assessed' if applicable.

#### 4.3.12 Intra-annotator agreement

**Question 4.3.12.1: How was the intra-annotator agreement between evaluators measured for this quality criterion?**

*What to enter in the text box:* The method(s) used for measuring intra-annotator agreement. If intra-annotator agreement was not measured, enter 'IntraAA not assessed'.

**Question 4.3.12.2: What was the intra-annotator agreement score?**

*What to enter in the text box:* The intra-annotator agreement score(s) obtained with the method(s) in Question 4.3.12.1. Enter 'IntraAA not assessed' if applicable.

#### HEDS Section 5: Ethics

**Question 5.1: Which research ethics committee has approved the evaluation experiment this sheet is being completed for, or the larger study it is part of?**

*What to enter in the text box:* Normally, 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 ethical approval obtained' if applicable.

**Question 5.2: Does personal data (as defined in GDPR Art. 4, §1: <https://gdpr.eu/article-4-definitions>) occur in any of the system outputs (or human-authored stand-ins) evaluated, or responses collected, in the experiment this sheet is being completed for?**

*Multiple-choice options (select one):*

- ☐ **No, personal data as defined by GDPR was neither evaluated nor collected.**
- ☐ **Yes, personal data as defined by GDPR was evaluated and/or collected:** Explain in the text

box, how it was ensured that the personal data was handled in accordance with GDPR.

**Question 5.3: Does special category information (as defined in GDPR Art. 9, §1: <https://gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited>) occur in any of the evaluation items evaluated, or responses collected, in the evaluation experiment this sheet is being completed for?**

*Multiple-choice options (select one):*

- ☐ **No, special category data as defined by GDPR was neither evaluated nor collected.**
- ☐ **Yes, special category data as defined by GDPR was evaluated and/or collected:** Explain in the text box how it was ensured that the special-category data was handled in accordance with GDPR.

**Question 5.4: Have any impact assessments been carried out for the evaluation experiment, and/or any data collected/evaluated in connection with it?**

*What to enter in the text box:* If an *ex ante* or *ex post* impact assessment has been carried out, and the assessment plan and process, as well as the outcomes, were captured in written form, describe them here and link to the report. Otherwise enter 'no impact assessment carried out'. Types of impact assessment include data protection impact assessments, e.g. under [GDPR](#). Environmental and social impact assessment frameworks are also available.

## 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, Simon Mille, and Craig Thomson. 2025a. Standard quality criteria derived from current nlp evaluations for guiding evaluation design and grounding comparability and ai compliance assessments. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.
- Anya Belz, Simon Mille, and Craig Thomson. 2025b. A taxonomy of quality criterion names and definitions for evaluating nlp systems in terms of standard comparable aspects of quality.
- Anya Belz, Simon Mille, Craig Thomson, and Rudali Huidrom. 2024. [QCET: An interactive taxonomy of quality criteria for comparable and repeatable evaluation of NLP systems](#). In *Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations*, pages 9–12, Tokyo, Japan. 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*.
- Anya Belz, Anastasia Shimorina, Maja Popovic, and Ehud Reiter. 2022. The 2022 reprogen shared task on reproducibility of evaluations in nlg: Overview and results. *INLG 2022*, page 43.
- Anya Belz and Craig Thomson. 2023. [The 2023 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results](#). In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, pages 35–48, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- Anya Belz and Craig Thomson. 2024. The 2024 reproNLP shared task on reproducibility of evaluations in nlp: Overview and results. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval)@ LREC-COLING 2024*, pages 91–105.
- Anya Belz, Craig Thomson, Javier González-Corbelle, and Malo Ruelle. 2025c. The 2025 reproNLP shared task on reproducibility of evaluations in nlp: Overview and results. In *Proceedings of the 4th Workshop on Generation, Evaluation & Metrics (GEM<sup>2</sup>)*.
- Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood,

- Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023a. [Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP](#). In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.
- Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023b. [Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP](#). In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.
- Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023c. [Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.
- 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.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. [Datasheets for datasets](#). *arXiv preprint arXiv:1803.09010*.
- Sebastian Gehrmann, Tosin Adewumi, Karmanya Agarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D. Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, 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, Rubungo Andre Niyongabo, 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 Cabezero, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. [The GEM benchmark: Natural language generation, its evaluation and metrics](#). *arXiv preprint arXiv:2102.01672*.
- 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.
- Karen Sparck Jones. 1981. *Information retrieval experiment*. Butterworth-Heinemann.
- H. Kamp and U. Reyle. 2013. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht.
- 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.
- 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:2103.09710*.
- Anastasia Shimorina and Anya Belz. 2022. [The human evaluation datasheet: A template for recording details of human evaluation experiments in NLP](#). In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 54–75, Dublin, Ireland. Association for Computational Linguistics.
- Craig Thomson and Anya Belz. 2024. [\(mostly\) automatic experiment execution for human evaluations of NLP systems](#). In *Proceedings of the 17th International Natural Language Generation Conference*, pages 272–279, Tokyo, Japan. Association for Computational Linguistics.
- Craig Thomson, Ehud Reiter, and Belz Anya. 2024. [Common flaws in running human evaluation experiments in nlp](#). *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.

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.