## Repro-HEDS: Property Names and Value Ranges for Recording the Experimental Properties of Human Evaluations of NLP Systems for Pre-Registration and Reproduction

*Authors: Craig Thomson and Anya Belz, last updated: 23 July 2023*

This document describes a set of property names and value ranges (adapted from HEDS, [Shimorina & Belz, 2022](https://aclanthology.org/2022.humeval-1.6.pdf)) for recording the experimental properties of human evaluation experiments in NLP, for the purposes of preregistration and forming part of publications and research repositories. The properties are designed to capture the minimum information about an experiment that is required to repeat the experiment, rather than an exhaustive list of properties that would fully define an experiment (as is the aim in HEDS).

For pre-registration, the experimental properties are all that is needed; experimental property names are highlighted in green below. For published work and reproduction experiments, we provide an additional set of properties, relating to the paper in which the original experiment was reported; the names of these properties are highlighted in yellow.

* Properties of the paper
* Properties of the experiment

### Information recorded for properties

The properties selected for inclusion in Repro-HEDS are those that must be known in order to be able to carry out a repetition of an experiment, at the level of detail at which there is a reasonable expectation of an effect on results. For example, an experiment that uses crowd-sourced evaluators might well yield different results than one that is in all respects the same but uses non-crowdsourced evaluators. However, two well-designed identical experiments that use two different crowd-sourcing platforms should not yield significantly different results.

The experimental properties included here are intended to serve as the conditions of measurement that are taken into account when performing quantified reproducibility assessments, or QRA ([Belz, 2022](https://direct.mit.edu/coli/article/48/4/1125/112113/A-Metrological-Perspective-on-Reproducibility-in)).

#### Property keys

Each property has a number of keys that define it; these are shown with an underline. These keys are always presented in the below order, although if their corresponding value would be blank, they are omitted for brevity. Additionally, we provide notes and reference, in italics as below.

Question: The question that should be answered by the property.

Subheadings: Subheadings are used where multiple values can be entered for one property. Each subheading will share all other keys of the parent property. Properties with subheadings are marked by an asterisk next to the property name below.

What to record: Describes what should be recorded for this property.

Fixed list: If the property must be selected from a list of categorical values, these values are listed here.

*Notes*: Additional notes for anything that is relevant to the property, or for any examples.

*References*: Links to papers or other resources that the property definition is based upon.

#### Marking experiment properties as unclear

When using Repro-HEDS in the context of a reproduction study, if the value of any property of the original experiment cannot be determined, the “unclear” value is used. For brevity, this “unclear” option is not included in individual property definitions below. Note that the “unclear” value does not apply to the corresponding reproduction experiment, or to preregistration, as in those cases, all values must be clear.

#### The definition of “an experiment”

For present purposes, an experiment is uniquely identified by the combination of values for the following properties:

* Dataset
* Languages
* Standardised Quality Criterion

Note that only single values are allowed for these three properties (unlike some of the other properties below). In the event that a paper reports multiple experiments sharing the same values for the above three properties, the experiments in question are recorded separately. This would happen, for example, if a paper reported a reproduction of an experiment within the same paper as the original experiment.

### Paper property definitions (not for pre-registration)

#### Venue

Question: Which venue was the paper published at?

What to record: The venue at which the paper was published, e.g., *ACL Main Conference* or *Journal of Artificial Intelligence Research*.

#### Paper URL

Question: What is the URL for the paper?

What to record: A link to the paper.

#### Title

Question: What is the Title of the paper?

What to record: The title of the paper.

#### Year

Question: Which year was the paper published?

What to record: The year of publication for the paper.

#### Resources URLs

Question: Where can information relating to the property values found?

What to record: A list of links to sources of information that can be used to verify the information recorded in the properties, e.g. supplementary material PDF or github repository.

### Experiment property definitions

#### Dataset

Question: Which dataset is used?

What to record: The name of the dataset used for this experiment.

*Notes:* Only a single value is allowed. If more than one dataset is used, create multiple records for experiments associated with each.

#### Languages

Question: Which language(s) are evaluated?

What to record: The language(s) that are evaluated in this experiment.

*Notes:* Only a single value is allowed, except where a single dataset incorporates multiple languages, e.g. in tasks such as machine translation, where a language pair in the form “XX to YY”, e.g., “EN to DE” should be recorded, using the language codes or long forms from ISO 639-1 (2019).[[1]](#footnote-0)

#### Task

Question: Which NLP task is evaluated?

What to record: A free-text description of the NLP task that is being evaluated.

*Notes*: Since NLP includes many tasks, some of which have fuzzy boundaries, it is outwith the scope of this document to define them all. However, note that the task should describe the mapping from inputs to outputs.

#### Control for Native and Non-native Speakers

Question: Does the experiment include control for recruiting only native speakers or only non-native speakers?

What to record: The appropriate option from the fixed list below.

Fixed list:

* **yes native** - The experiment includes checks that participants are native speakers of the language being evaluated (i.e. more than just regional recruitment as below).
* **yes non-native** - The experiment includes checks that participants are non-native speakers of the language being evaluated (more than regional recruitment).
* **yes mix** - A mix of native and non-native speakers are purposefully recruited for the experiment. Note that if results are reported for these groups separately, rather than as an ensemble, this would be considered two experiments.
* **region** - Participants from a region where the language is natively spoken are recruited (no other checks included).
* **no** - There is no control for the native vs. non-native status of participants.

#### Participant Type \*

Question: What types of participants are recruited?

Subheadings: Record multiple values as needed to describe all participant type(s) recruited.

What to record: The appropriate option(s) from the fixed list below.

Fixed list:

* **Undergrad / Masters Students**
* **Other researchers / colleagues** -this includes PhD students
* **Other specific type** - identify type (but not expertise which is covered below)
* **No specific type** - No specific type of participant is recruited, e.g. workers on a crowd platform or participants from social media

#### Author Involvement

Question: Are participants in the experiment also authors on the paper?

What to record: The appropriate option from the fixed list:

Fixed list:

* **All** - All participants are authors
* **Some** - Some participants are authors
* **None** - No participants are authors

#### Intrinsic or Extrinsic

Question: Is the evaluation intrinsic or extrinsic?

What to record: The appropriate option from the fixed list:

Fixed list:

* **extrinsic** - The 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.
* **intrinsic** - All other cases.

*References*: [Belz et al. (2020)](https://aclanthology.org/2020.inlg-1.24/)

#### Absolute or Relative

Question: Is the evaluation absolute or relative?

What to record: The appropriate option from the fixed list:

Fixed list:

* **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.

*References*: [Belz et al. (2020)](https://aclanthology.org/2020.inlg-1.24/)

#### Objective or Subjective

Question: Is the evaluation objective or subjective?

What to record: The appropriate option from the fixed list below.

Fixed list:

* **objective** - The evaluation uses objective assessment, e.g. any automatically counted or otherwise quantified measurements such as mouse-clicks, occurrences in text, etc.
* **subjective** - All other cases. Subjective assessments involve ratings, opinions and preferences by evaluators.

*References*: [Belz et al. (2020)](https://aclanthology.org/2020.inlg-1.24/)

#### Crowd Platform

Question: Is a crowd platform used?

What to record: The appropriate option from the fixed list below.

Fixed list:

1. **Yes - public**: A publicly available crowd platform such as Amazon Mechanical Turk or Prolific.
2. **Yes - private**: A private crowd platform similar to the above, but only accessible internally within an organisation.
3. **No**

#### Expertise Sought

Question: Is expertise sought?

What to record: The appropriate option from the fixed list below.

Fixed list:

* **Yes** - all participants are required to have one or more types of expertise.
* **No** - participants are not required to have specific expertise.

*Notes*: Do not consider participants' (non)native speaker status as expertise.

#### Expertise Types

Question: If expertise is sought, what type?

What to record: Details of the type(s) of expertise that participants are required to have. If no expertise is sought, enter “N/A.”

#### Total Items

Question: How many items are there in total?

What to record: The number of items that are evaluated.

*Notes*: The definition of an item is:

For **absolute** evaluations: One system output that is judged absolutely by participants for one quality criterion.

For **relative** evaluations: A non-singleton set of system outputs that are judged relatively by participants for one quality criterion.

#### Total Systems

Question: How many systems are evaluated?

What to record: The number (integer) of systems that are evaluated as part of the experiment. Include in the count any reference outputs, as well as any baselines, that are evaluated alongside systems.

#### Total Participants

Question: What is the controlled number of total participants?

What to record: The appropriate option from the fixed list below.

Fixed list:

* **INT**: An integer value for the total number of participants.
* **varies [unbounded]**: When each participant is shown all items, but the total number of participants is “as many as can be found”.
* **varies [uncontrolled]**: When neither of the above can describe the total number of participants.

*Notes*: The total number of participants is inherently limited to the total number of judgments required. “varies [uncontrolled]” can be used with this caveat.

#### Participants per Item

Question: How many participants per item are there?

What to record: The appropriate option from the fixed list below.

Fixed list:

* **INT**:An integer value for the number of items judged by each participant.
* **varies [assignment known]**: When each item is assigned a known number of participants, but this number varied between items.
* **varies [unbounded]**: When each participant is shown all items, but the total number of participants is “as many as can be found”.
* **varies [uncontrolled]**: When participants are free to choose which items they judge.

*Notes*: The total number of participants per item is inherently limited by the total number of items (if known). “varies [uncontrolled]” can be used with this caveat.

#### Items per Participant

Question: How many items per participant are there?

What to record: The appropriate option from the fixed list below.

Fixed list:

* **INT**: An integer value for the number of items judged by each participant.
* **varies [assignment known]**: When the number of items judged by each participant is known but differs between participants.
* **varies [limited]**: When participants are restricted in how many items they could judge, for example, “no more than 20”, or “no fewer than 5”, but otherwise there is no control.
* **varies [uncontrolled]**: When none of the above can describe the total number of participants per item.

*Notes*: The total number of items per participant is inherently limited to the total number of items. “varies [uncontrolled]” can be used with this caveat.

#### Verbatim Quality Criterion

Question: What is the criterion name and definition that is used and where can it be found?

What to record: The verbatim quality criterion label and corresponding definition (if present) that is used.

*Notes:* In determiningthe value of this property, the paper, evaluation interface and any instructions given to evaluators should be taken into account. If the information provided in these three sources doesn’t align completely, priority should be given to what it says in the evaluation interface, followed by instructions.

#### Standardised Quality Criterion \*

Question: Which criterion or criteria from the taxonomy does the verbatim quality criterion most closely correspond to?

Subheadings: Select as many criteria as necessary to cover the verbatim quality criterion**.**

What to record: Any of the criterion names from the QCET Taxonomy of normalised quality criteria (based on that in [Howcroft et al. 2020](https://aclanthology.org/2020.inlg-1.23/)) that either fully or partially correspond to the verbatim quality criterion.

*Notes*: For example, if the verbatim quality criterion is: “**fluency**: whether the text is grammatical and coherent” then entries would be recorded under three subheadings, one for fluency, one for grammaticality, and one for coherence.

*References*: [Howcroft et al. 2020](https://aclanthology.org/2020.inlg-1.23/)

#### Rating Instrument Type \*

Question: What are the types of rating instruments shown to participants to evaluate each item?

Subheadings: Under each subheading, record a type of instrument that is used to judge the item as part of a single judgement.

What to record: Under each subheading, an option from the fixed list.

Fixed list:

* Free text fields for everything else
* Free text fields for single, item-level judgments
* Integer count of occurrences in output (without marking spans)
* Multiple choice select multiple (clicking)
* Multiple choice select one (any other method)
* Multiple choice select one (clicking)
* Multiple choice select one (user types label from a list)
* Ordering a set (any other method)
* Ordering a set (characters/integers)
* Ordering a set (drag and drop)
* Select word span boundary (optionally with label)
* Single character/integer input for single, item-level judgments
* Slider

*Notes*: In most cases, there will only be one instrument type that is used to evaluate a single judgement, although there may be multiple instruments, for example, when:

* participants are asked to select a multiple-choice option and provide free-text feedback.
* participants are asked to judge (multiple choice; yes/no) whether an output has any problems, and then record the severity of those problems with a separate instrument (integer input).

#### Outputs per Judgement

Question: How many system outputs (or parts) are considered in a single judgement?

What to record: How many system outputs (or parts) are shown to participants within a single item (see “Total Items” for the definition of “an item”.

Fixed list:

* Single
* Pair
* More

*Notes*: “Parts” here refers to situations where parts of the system output are the atomic items that are evaluated (rather than the whole output from the perspective of the model). For example, if a data-to-text system outputs a paragraph, an experiment may be set up where sentences from generated paragraphs are evaluated, rather than the whole output.

### Guidance on recording properties

The aim of the present document is to provide Repro-HEDS property definitions. We have created a Google spreadsheet template that can be used to record Repro-HEDS properties. A companion document explains how to complete this template.

Note that we recommend that whatever format the properties are recorded in, the source where a property value can be verified should always be recorded.

### 

1. <https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes> [↑](#footnote-ref-0)