Human Evaluation of NLP System Quality

INLG Tutorial, 24th September 2024

Unit 4: Experiment Design

Link to Unit 4 Resources

Overview

Unit 4: Experiment Design

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Design decisions and experiment properties
3. Step 3: Output sampling and normalisation
4. Step 4: Rating instrument
5. Step 5: Type and characteristics of evaluators
6. Step 6: Evaluator recruitment and training
7. Step 7: Conditions under which evaluators carry out experiment
8. Step 8: Quality assurance methods
9. Steps 10 and 11: Impact assessment and ethical review
10. Completing human-evaluation datasheet (HEDS)
11. Unit summary and pointers to other units
12. References

Overview

Unit 4: Experiment Design

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Design decisions and experiment properties
3. Step 3: Output sampling and normalisation
4. Step 4: Rating instrument
5. Step 5: Type and characteristics of evaluators
6. Step 6: Evaluator recruitment and training
7. Step 7: Conditions under which evaluators carry out experiment
8. Step 8: Quality assurance methods
9. Steps 10 and 11: Impact assessment and ethical review
10. Completing human-evaluation datasheet (HEDS)
11. Unit summary and pointers to other units
12. References

Unit aims and learning outcomes

- The aims of Unit 4 are:
  - To take a closer look at the remaining steps in Phase I (Design).
  - To present design options for each step and consider their suitability in different evaluation contexts.
  - To introduce the Human Evaluation Data Sheet (HEDS) for capturing details of an experiment e.g. for preregistration.
- After completion of the unit, participants will know:
  - What the different elements are that need to be specified in experiment design for human evaluations.
  - How to make informed choices from the range of options available for each element.
  - Create a complete design for a human evaluation that is suitable for a given evaluation measure and evaluation context.

Prerequisites and connections with other units

- Prerequisite(s) of Unit 4: Units 2 and 3. In particular Unit 3 and the taxonomy of quality criteria and evaluation modes introduced there.
- Unit 4 is a prerequisite of Units 5–8, as it introduces design choices used in later units.

Overview

Unit 4: Experiment Design

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Design decisions and experiment properties
3. Step 3: Output sampling and normalisation
4. Step 4: Rating instrument
5. Step 5: Type and characteristics of evaluators
6. Step 6: Evaluator recruitment and training
7. Step 7: Conditions under which evaluators carry out experiment
8. Step 8: Quality assurance methods
9. Steps 10 and 11: Impact assessment and ethical review
10. Completing human-evaluation datasheet (HEDS)
11. Unit summary and pointers to other units
12. References

From the ‘what’ to the ‘how’

- Recall from Unit 2:
  Quality criterion + evaluation modes = **evaluation measure**;
  Evaluation measure + experimental design = **evaluation method**.
- In Unit 3, we got as far as having fully specified evaluation measures *m* .
- Can think of *m* as ***what*** is being evaluated.
- In Unit 4, we look at creating the experiment design that’s needed for an evaluation method E_m that assesses a given evaluation measure *m.*
- Can think of this as specifying ***how*** we evaluate *m.*
- Experiment design can be broken down into different aspects which we call **experiment (design) properties**.
- Will provide an overview of all experiment properties first via the HEDS questions corresponding to them.
- Then go through the 11 steps via which experiment properties are specified.

Experiment design properties recorded in HEDS

3.1.2: How are system outputs (or other evaluation items) selected for inclusion in the evaluation experiment?
Multiple choice or ‘other’ with description.

3.1.1: How many system outputs (or other evaluation items) are evaluated per system in the evaluation experiment?
Integer.

3.1.3.1: What method was used to determine the statistical power of the sample size?
Name of method used.

3.1.3.2: What is the statistical power of the sample size?

Numerical results of statistical power calculation on the sample.

3.2.1: How many evaluators are there in this experiment?
Total number of evaluators participating in the experiment.

Colour coding maps to the 11 steps in specifying experiment properties.

Experiment design properties recorded in HEDS

3.2.2.1: What kind of evaluators are in this experiment?
Multiple choice or ‘N/A’ with explanation.

3.2.2.2: Were the participants paid or unpaid?
Multiple choice or ‘N/A’ with explanation.

3.2.2.3: Were the participants previously known to the authors?
Multiple choice or ‘N/A’ with explanation.

3.2.2.4: Were one or more of the authors among the participants?
Multiple choice or ‘N/A’ with explanation.

3.2.2.5: Further details for participant type.
Text elaborating on selections for questions 3.2.2.1 to 3.2.2.4 above.

3.2.5: What other characteristics do the evaluators have (as per qualifying criteria, or information gathered in evaluation).
Text listing any characteristics not covered in previous questions that evaluators are known to have.

Experiment design properties recorded in HEDS

3.2.3: How are evaluators recruited?
Text describing recruitment process in detail.

3.2.4: What training and/or practice are evaluators given before starting on the evaluation itself?
Text explaining training/practice including any introductory explanations, e.g. in evaluation interface.

3.3.6: Are evaluators told they can ask questions about the evaluation and/or provide feedback?
Check-box list.

4.3.1: What do you call the quality criterion in explanations/interfaces to evaluators?
Text giving the QC name, e.g. selected from QCET taxonomy; ‘N/A’ if none used.

4.3.2: What definition do you give for the quality criterion in explanations/interfaces to evaluators?
Text giving verbatim definition or question given to evaluators; ‘N/A’ if none used.

3.3.3.1: What quality assurance methods are used to ensure evaluators and/or their responses are suitable?
Check-box list; ‘other’ and text box for other methods; alternatively ‘none of the above’.

3.3.3.2: Please describe in detail the quality assurance methods that were used.
Text describing the methods in 3.3.3.2, or ‘N/A’.

Experiment design properties recorded in HEDS

3.3.5: How free are evaluators regarding when and how quickly to carry out evaluations?
Check-box list.

3.3.7: What are the experimental conditions in which evaluators carry out the evaluations?
Multiple choice or ‘other’ with description.

3.3.8: Briefly describe the (range of different) conditions in which evaluators carry out the evaluations.
Text describing aspects not covered by 3.3.7, or ‘N/A’.

Experiment design properties recorded in HEDS

3.3.2: How are responses collected?
Text describing process used to collect responses, e.g. entering text in a forms.

3.3.4.1: Please include a link to online copies of the form/interface that was shown to participants.
Text giving URL.

3.3.4.2: What do evaluators see when carrying out evaluations?
Text further describing what evaluators are shown, e.g. dynamic or changing elements.

4.3.3: Are the rating instrument response values discrete or continuous? If so, please also indicate the size.
Multiple choice or ‘N/A’ with explanation.

4.3.4: List or range of possible values of the scale or other rating instrument.
List or range of possible response values (of size specified in Question 4.3.3). ‘N/A’ if no rating instrument is used.

4.3.5: How is the scale or other rating instrument presented to evaluators?
Multiple choice or ‘other’ with description.

Experiment design properties recorded in HEDS

4.3.6: If there is no rating instrument, what task do evaluators perform, and what information is recorded?
Text describing the task evaluators perform if there is no rating instrument; ‘N/A’ if there *is* a rating instrument.

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

4.3.8: Form of response elicitation (e.g. relative quality assessment).
Multiple choice or ‘Other’ with description.

4.3.9: How are raw responses from participants aggregated or otherwise processed to obtain reported scores?
Text describing method(s) used in the conversion(s) or ‘N/A’ with explanation.

4.3.10: Method(s) used for determining effect size and significance of findings for this quality criterion.
List of method(s); or ‘None’ if none used.

Design choices and experiment design properties

The 11 steps in making design choices and how they map to experiment properties:

1. From steps covered in Unit 3 we have
   - ○ **Quality criteria** and **evaluation modes** → HEDS Section 4, Questions 4.1.1–4.2.3
   - ○ Set of **systems** we want to evaluate → HEDS Section 2 (Questions re systems evaluated)
2. **Number of outputs and evaluators**: Perform power calculations to determine sample size (numbers of outputs/evaluators and how assigned) → HEDS 3.1.1, 3.1.3.1, 3.1.3.2, 3.2.1
3. Choose **output sampling method** → HEDS 3.1.2 (note that we need the implementation from Phase II before we can complete 3.1.3.3 re code)
4. **Rating instrument**: type, range of response values, how presented to evaluators, how evaluators interact with instrument → HEDS 3.3.2, 3.3.4.1, 3.3.4.2, 4.3.1, 4.3.3–4.3.8
5. Decide type and (qualifying) characteristics of **evaluators** → HEDS 3.2.2.1–3.2.3, 3.2.5

Design choices and experiment design properties

The 11 steps in making design choices and how they map to experiment properties (cont.):

6. **Evaluator recruitment, training, instructions** → HEDS 3.2.3, 3.2.4, 3.3.6, 4.3.1, 4.3.2
7. **Conditions** under which evaluators carry out experiment → HEDS 3.3.5, 3.3.7, 3.3.8
8. Select/design **quality assurance** methods → HEDS 3.3.3.1, 3.3.3.2
9. Select methods for **analysis of results** → HEDS 4.3.9, 4.3.10
10. Carry out **impact assessment** → HEDS 5.4 (see later section)
11. Conduct **ethical review** → HEDS 5.1–5.3 (see later section)

Where covered in tutorial:

- Step 1 was covered in Unit 3.
- Steps 2 and 9 will be covered in Unit 5.
- In the next sections in this unit, will look at each of Steps 3–8, 10 and 11 in turn.

Overview

Unit 4: Experiment Design

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Design decisions and experiment properties
3. Step 3: Output sampling and normalisation
4. Step 4: Rating instrument
5. Step 5: Type and characteristics of evaluators
6. Step 6: Evaluator recruitment and training
7. Step 7: Conditions under which evaluators carry out experiment
8. Step 8: Quality assurance methods
9. Steps 10 and 11: Impact assessment and ethical review
10. Completing human-evaluation datasheet (HEDS)
11. Unit summary and pointers to other units
12. References

Step 3: Output sampling

Context:

- Once the systems to be evaluated have been selected, samples of their behaviour need to be obtained for presentation to evaluators.
- This usually means generating system outputs for the same set of inputs, but can be:
    - A sequence of user and system turns (as in dialogue tasks);
    - Other user-system interactions, etc.

Considerations:

- A large enough sample needs to be obtained for the desired statistical power of the experiment – determined in Step 2 of Phase I (Design), see Unit 5.
- Inputs need to be selected so their characteristics are representative of the system task as a whole, e.g. by stratified random sampling.

Step 3: Output sampling

Task:

- To write **specifications for output sampling**, mapping from systems to directly comparable samples of their behaviour that are representative of the system task as a whole. Ready to be implemented in Phase II.

Example:

- For WebNLG 2023, the shared task test set was randomly sampled with stratification for:
  - Number of input triples, e.g.: `region (Arròs_negre, Valencian_Community); ingredient (Arròs_negre, Cuttlefish)`
  - Categories of triples, e.g.: `Food`
- The script took (i) the test set inputs, and (ii) the corresponding system outputs as input, and produced as output a data structure with parallel samples of system outputs for the same subset of inputs, selected to be as representative as possible in the above terms.

Step 3: Output normalisation

Context:

- System outputs that have been collected from evaluated systems may differ in ways that we don’t want to affect assessment, e.g. differences in spacing and capitalisation.

Considerations:

- Need to be careful to distinguish between elements that we do wish to affect assessment and those we do not.
- E.g. whether a system leaves in HTML tags probably should be taken to be a reflection of its performance and not be changed.
- However, whether a system follows the convention of leaving two space characters after sentence-ending punctuation should arguably not be taken into account.
- The aim is to create a level playing field that is fair to all systems evaluated.

Step 3: Output normalisation

Task:

- To write **specifications for output normalisation**, mapping from system outputs as produced by systems to normalised character sequences according to a specified set of normalisation rules.

Examples:

- Multilingual Surface Realisation Shared Task (Mille et al., 2018): Output texts were normalised prior to computing metrics by lower-casing all tokens, removing any extraneous whitespace characters.
- WebNLG 2023 Shared Task (Cripwell et al., 2023): No text normalisation, but inputs were mapped to more human-readable, bracketed format.

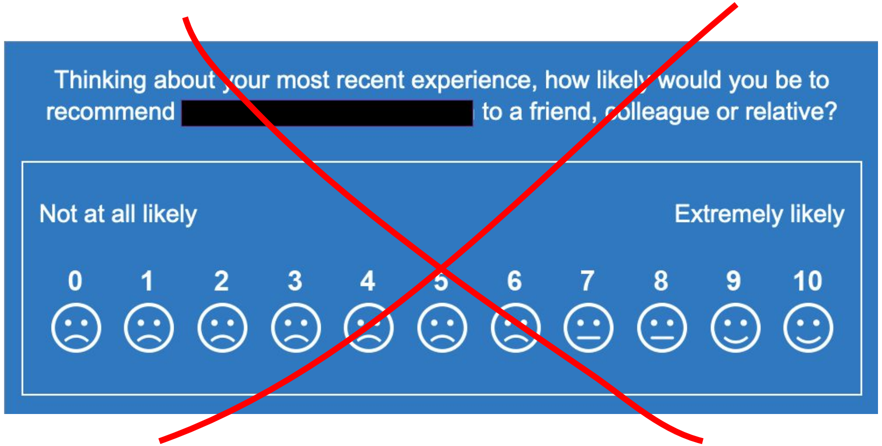
Overview

Unit 4: Experiment Design

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Design decisions and experiment properties
3. Step 3: Output sampling and normalisation
4. Step 4: Rating instrument
5. Step 5: Type and characteristics of evaluators
6. Step 6: Evaluator recruitment and training
7. Step 7: Conditions under which evaluators carry out experiment
8. Step 8: Quality assurance methods
9. Steps 10 and 11: Impact assessment and ethical review
10. Completing human-evaluation datasheet (HEDS)
11. Unit summary and pointers to other units
12. References

Step 4: Rating Instrument

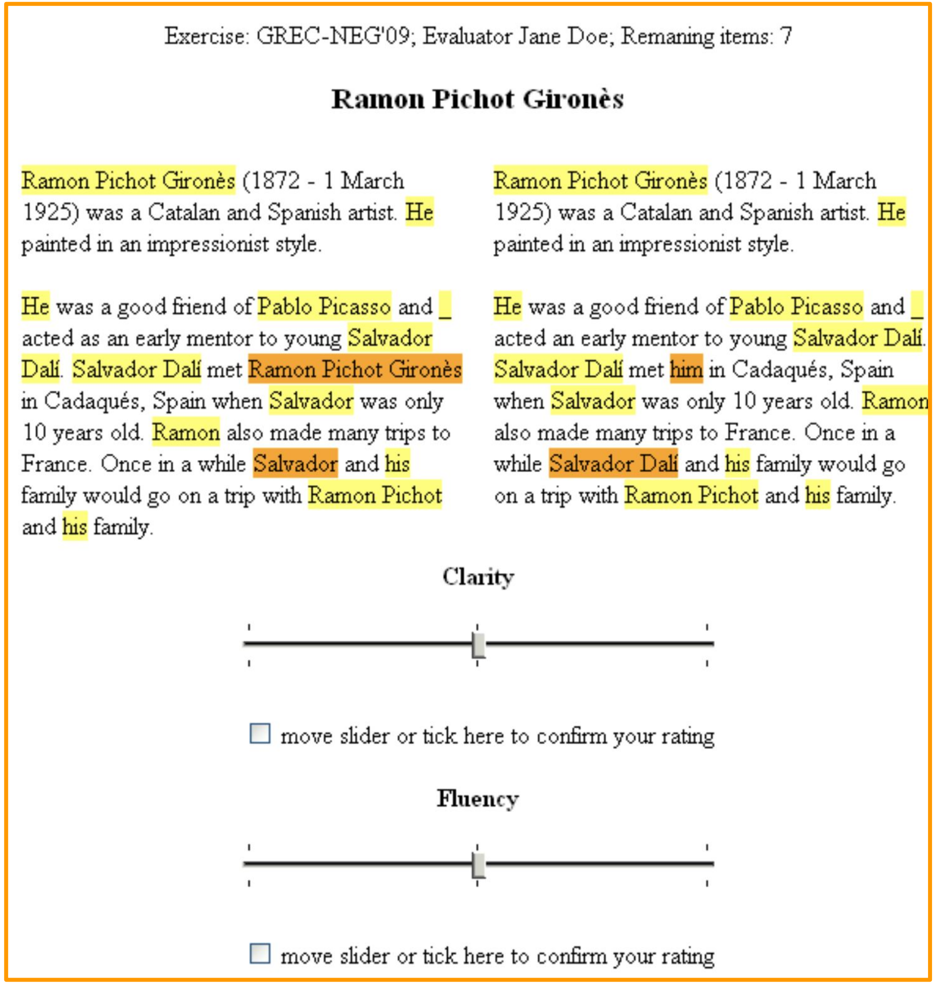
- **Context**:
  - Need to find a way for evaluators to provide responses (measured values v_i) for the evaluation measure.
  - This can e.g. take the form of a rating scale or a list to be ranked.
- **Considerations**:
  - Rating instruments should be easy to use, the risk of making mistakes low.
  - Depending on what the data type of the collected responses is, different forms of aggregation and analysis are appropriate.
  - Verbal or pictorial labels on scales can make a big difference, so use with care, or not at all.
  - How the rating instrument is presented also matters: interface should be clear, easy to absorb, and include essential instructions.



Thinking about your most recent experience, how likely would you be to recommend [redacted] to a friend, colleague or relative?

Step 4: Rating Instrument

- **Task**: To write a **specification identifying the selected rating instrument**, including its size/range, data type, visual appearance, labels, evaluator interaction, and surrounding interface.
- **Example**:
  - GREC 2009 shared task evaluation measures:
    - *Relative, Subjective, Intrinsic Referential Clarity*
    - *Relative, Subjective, Intrinsic Fluency*
  - Range: -10.0 to +10.0 (ordinal, 0.1 step size)



Exercise: GREC-NEG'09; Evaluator Jane Doe; Remaning items: 7

Ramon Pichot Gironès

Ramon Pichot Gironès (1872 - 1 March 1925) was a Catalan and Spanish artist. He painted in an impressionist style.

He was a good friend of Pablo Picasso and _ acted as an early mentor to young Salvador Dalí. Salvador Dalí met Ramon Pichot Gironès in Cadaqués, Spain when Salvador was only 10 years old. Ramon also made many trips to France. Once in a while Salvador and his family would go on a trip with Ramon Pichot and his family.

Ramon Pichot Gironès (1872 - 1 March 1925) was a Catalan and Spanish artist. He painted in an impressionist style.

He was a good friend of Pablo Picasso and _ acted an early mentor to young Salvador Dalí. Salvador Dalí met him in Cadaqués, Spain when Salvador was only 10 years old. Ramon also made many trips to France. Once in a while Salvador Dalí and his family would go on a trip with Ramon Pichot and his family.

Clarity

□ move slider or tick here to confirm your rating

Fluency

□ move slider or tick here to confirm your rating

Overview

Unit 4: Experiment Design

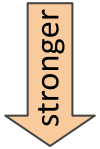
1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Design decisions and experiment properties
3. Step 3: Output sampling and normalisation
4. Step 4: Rating instrument
5. Step 5: Type and characteristics of evaluators
6. Step 6: Evaluator recruitment and training
7. Step 7: Conditions under which evaluators carry out experiment
8. Step 8: Quality assurance methods
9. Steps 10 and 11: Impact assessment and ethical review
10. Completing human-evaluation datasheet (HEDS)
11. Unit summary and pointers to other units
12. References

Step 5: Type and characteristics of evaluators

- **Context**:
  - Evaluators’ expertise and experience needs to be matched to the evaluation task.
  - This may just mean they are of the intended target user type.
  - Or it may mean hiring domain experts, e.g. translators to evaluate MT systems, and doctors to evaluate consultation note generators (Moramarco et al., 2022).
- **Considerations**:
  - Generally, the more qualified evaluators are, the more expensive they are.
  - Also the more difficult they may be to get hold of.
  - Some results show high correlation between expert and layperson evaluations in some domains, e.g. weather forecast text generation (Reiter & Belz, 2006).
  - Consider carefully what knowledge/expertise people need to have in order to carry out your evaluation tasks. Examples: levels of numeracy, literacy, fluency in specific language(s), domain knowledge, etc.
  - Can you provide them with that knowledge/expertise in training or is that not feasible?

Step 5: Type and characteristics of evaluators

- **Considerations (cont.)**:
  - Whatever levels of knowledge/expertise you decide is appropriate for the evaluation task, participants will need to be screened for having those levels.
  - This can be done through three main options:
    - Self-attested screening
    - Evidence-based screening
    - Screening tests of knowledge/expertise
  - In addition, you may wish to ask participants to perform the evaluation task on a small test sample (knowledge/expertise is not always enough, Cripwell et al., 2023)



stronger

- **Task**: Write an **evaluator type/characteristics specification**, detailing:
  - The characteristics that participants need to have,
  - A regime for screening participants for the characteristics, and
  - Qualification exercise(s) participants need to carry out.

Step 5: Type and characteristics of evaluators

- **Examples**:
  1. For WebNLG 2023, the Evaluator type/characteristics protocol specified the following:
     - Participant characteristics: had to be experienced professional translators.
     - Screening participants for the characteristics: recruitment through a translation agency, confirming qualification for the task.
     - Qualification exercise: perform the task on a test sample of 10 input/output pairs; had to achieve minimum 0.7 Fleiss's kappa inter-annotator agreement with authors' annotations; one round of feedback and try again allowed.
  2. Thomson & Reiter (2020):
     - Participant characteristics: Basketball domain knowledge, literacy, numeracy.
     - Screening participants for the characteristics: MTurk Masters, US bachelor's degree (self-reported) & domain knowledge (self-reported, regularly watches or plays).
     - Qualification exercise: Participants have to get 70% correct (compared to gold standard annotations) on a set of qualification evaluation items.

Overview

Unit 4: Experiment Design

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Design decisions and experiment properties
3. Step 3: Output sampling and normalisation
4. Step 4: Rating instrument
5. Step 5: Type and characteristics of evaluators
6. Step 6: Evaluator recruitment and training
7. Step 7: Conditions under which evaluators carry out experiment
8. Step 8: Quality assurance methods
9. Steps 10 and 11: Impact assessment and ethical review
10. Completing human-evaluation datasheet (HEDS)
11. Unit summary and pointers to other units
12. References

Step 6: Evaluator recruitment and training

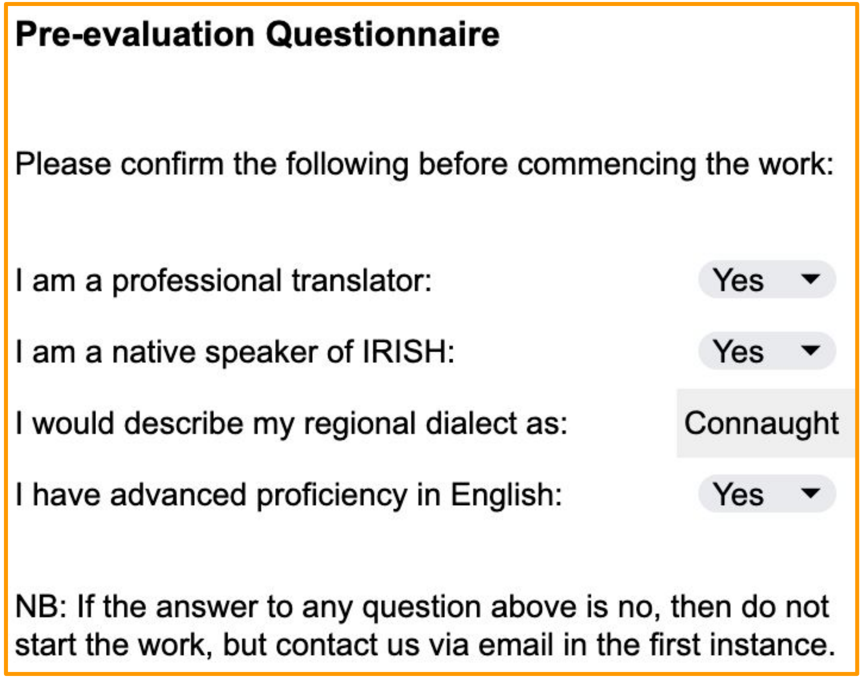
- **Context**: Type/characteristics protocol was specified in Step 5, now need to consider how to get hold of evaluators and what training they need to complement screened characteristics/abilities.
- **Considerations**:
    - It may be convenient to recruit students or colleagues in the same organisation, but will result in a biased sample, not representative of the wider population.
    - E.g. participants recruited from the NLP lab would have:
        - Knowledge of NLP systems and evaluation processes.
        - A certain level of education.
        - Possibly a certain level of language proficiency.
    - Can be an issue even for pilot experiments, because inter-annotator agreement may differ.
    - Some crowd platforms, such as Prolific, allow for:
        - Representative sampling
        - Quota sampling

Step 6: Evaluator recruitment and training

- **Considerations (cont.)**:
    - For recruitment, ensure that any adverts/posts reach a representative sample of the target population, e.g. multiple mailing lists, social media sites, universities in multiple geographic locations, as appropriate for the evaluation task.
    - It's usually a good idea to provide training and instructions before the evaluation to homogenise how evaluators perform their task.
    - It's hard to write clear training/instruction texts that evaluators will interpret similarly – example assessments can provide more clarity.
- **Task**: Write an **evaluator recruitment/training protocol** specifying:
    - The text and other information included in recruitment communication.
    - The addressees and/or mailing lists, online platforms, public events, etc. where recruitment communication will be shared.
    - The training/instructions evaluators receive before the evaluation.

Step 6: Evaluator recruitment and training

- **Example**: WebNLG 2023:
  - Established translation agencies were contacted and asked to recommend translators for the four languages needed (Irish, Breton, Maltese, Welsh).
  - The recommended translators were then sent an email with the recruitment text and a link to a Google spreadsheet containing (i) self-attested screening questions about their expertise/knowledge (see right); and (ii) the qualification exercise (see Step 5 Example) with instructions.



Pre-evaluation Questionnaire

Please confirm the following before commencing the work:

I am a professional translator:	Yes ▾
I am a native speaker of IRISH:	Yes ▾
I would describe my regional dialect as:	Connaught
I have advanced proficiency in English:	Yes ▾

NB: If the answer to any question above is no, then do not start the work, but contact us via email in the first instance.

Resources for evaluator recruitment and training

- Writing training and instruction documents:
  - de Leeuw et al. (eds.) (2008) International Handbook of Survey Methodology.
  - Ruan et al. (2024) Defining and Detecting Vulnerability in Human Evaluation Guidelines: A Preliminary Study Towards Reliable NLG Evaluation.

Overview

Unit 4: Experiment Design

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Design decisions and experiment properties
3. Step 3: Output sampling and normalisation
4. Step 4: Rating instrument
5. Step 5: Type and characteristics of evaluators
6. Step 6: Evaluator recruitment and training
7. Step 7: Conditions under which evaluators carry out experiment
8. Step 8: Quality assurance methods
9. Steps 10 and 11: Impact assessment and ethical review
10. Completing human-evaluation datasheet (HEDS)
11. Unit summary and pointers to other units
12. References

Step 7: Conditions under which evaluators carry out experiment

- **Context**: Need to decide the degree to which we control where, in what time, etc., evaluators complete their part of the experiment.
- **Considerations**:
  - One reason for exerting more control is to achieve more similar conditions under which evaluators work (potentially leading to overall higher IAA).
  - Examples of controlled aspects:
    - Place: Anywhere they choose? In the lab? At a specific venue?
    - Time: Whenever they choose? Is there a time limit? All in one sitting?
    - Supervision: Will anyone be in the lab/room with them? Supervise/monitor remotely? Are they doing the experiment online? Can they ask questions?

Step 7: Conditions under which evaluators carry out experiment

- **Task**: Write a **controlled conditions protocol** specifying place, time, and any supervision you plan to carry out. (NB: attention checks and other monitoring come under quality assurance.)
- **Examples**:
  - WebNLG 2023: evaluators were free to carry out the evaluations at a place and time of their own choosing, via the online interface; they could ask questions and request clarifications at any point.
  - TUNA 2009 shared task (Gatt et al., 2009): for the extrinsic evaluation, evaluators performed the evaluation on the researcher’s laptop, in a quiet room (often their office), with the researcher present, available for questions/clarifications before and after, but not during assessments (as time measurements were being taken).

Overview

Unit 4: Experiment Design

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Design decisions and experiment properties
3. Step 3: Output sampling and normalisation
4. Step 4: Rating instrument
5. Step 5: Type and characteristics of evaluators
6. Step 6: Evaluator recruitment and training
7. Step 7: Conditions under which evaluators carry out experiment
8. Step 8: Quality assurance methods
9. Steps 10 and 11: Impact assessment and ethical review
10. Completing human-evaluation datasheet (HEDS)
11. Unit summary and pointers to other units
12. References

Step 8: Quality assurance (QA)

- **Context**:
  - Over and above screening for required characteristics and qualifying exercises, it can be useful to perform checks that evaluators are carrying out evaluations conscientiously and with sufficient quality.
  - The experiment design itself needs to be tested for robustness and reliability.
- **Considerations**:
  - Exclusion of only some items by an evaluator is sometimes seen in the literature, but hard to justify as a rigorous approach.
  - Forms of ad hoc, on-the-fly exclusion e.g. resulting from manual monitoring are also sometimes seen, but never a good idea.
  - Types of QA include attention checks, sanity checks, measuring time taken for assessments, post-evaluation correctness/quality checks where this is possible, etc.
  - If evaluators are caught not doing evaluations conscientiously or well enough, they will still need to be paid for work already done.

Step 8: Quality assurance (QA)

- **Considerations (cont.)**:
  - Failed checks will result in having fewer responses than planned and needing to recruit more evaluators (Thomson & Belz, 2024).
  - Interfaces and processes need to be tested for robustness and reliability, e.g.:
    - Can evaluators accidentally enter out-of-range responses?
    - Is the inter-annotator agreement high enough?
  - It is almost always the case that pilots and interface robustness tests lead to changes in the details of the design.
- **Task**: Write a **quality assurance protocol** detailing all measures as well specifications for code implementing (i) any attention checks and other forms of monitoring; (ii) all pre-final executions of the experiment planned for testing purposes; and (iii) any other tests you plan to carry out.

Step 8: Quality assurance (QA)

- **Example**: In the WMT 2019 (Barrault et al., 2019) and SR 2019 (Mille et al., 2019) shared tasks, system outputs were randomly assigned to MTurk HITs of 100 items, 20 of which were used solely for QA (i.e. they did not count towards system scores):

  i. Some evaluation items (outputs) were repeated as-is,
  ii. Some were repeated in a ‘damaged’ version, and
  iii. Some were replaced by their corresponding reference texts and repeated.

  In each case, a minimum threshold had to be reached for the HIT to be accepted:

  - for (i), scores must be similar enough,
  - for (ii) the score for the damaged version must be worse, and
  - for (iii) the score for the reference text must be higher.

  Up to 70% of HITs were discarded because evaluators failed these checks!

Step 8: Quality assurance (QA)

- **Example**: In WebNLG 2023, a pilot run revealed such low IAA for an originally planned *Absence of Substitutions* evaluation method, even among the researchers, that it was abandoned altogether.

Overview

Unit 4: Experiment Design

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Design decisions and experiment properties
3. Step 3: Output sampling and normalisation
4. Step 4: Rating instrument
5. Step 5: Type and characteristics of evaluators
6. Step 6: Evaluator recruitment and training
7. Step 7: Conditions under which evaluators carry out experiment
8. Step 8: Quality assurance methods
9. Steps 10 and 11: Impact assessment and ethical review
10. Completing human-evaluation datasheet (HEDS)
11. Unit summary and pointers to other units
12. References

Ethical review and impact assessment in HEDS

Ethical review and impact assessment are covered by HEDS Section 5:

- **5.1**: Has the evaluation experiment this sheet is being completed for, or the larger study it is part of, been approved by a research ethics committee?
  If yes, which research ethics committee?
- **5.2**: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain personal data (as defined in GDPR Art. 4, §1: https://gdpr.eu/article-4-definitions/)?
  If yes, describe data and state how addressed.
- **5.3**: Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain special category information (as defined in GDPR Art. 9, §1: https://gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited/)
  If yes, describe data and state how addressed.
- **5.4**: Have any impact assessments been carried out for the evaluation experiment, and/or any data collected/evaluated in connection with it?
  If yes, summarise approach(es) and outcomes.

Step 10: Impact assessment

Context:

- An impact assessment (IA) considers longer term and more indirect consequences of carrying out an evaluation experiment than ethical review normally covers.
- However, there is no broadly accepted standard for impact assessment for evaluation experiments.

Considerations:

- Few institutions currently require impact assessment for system evaluation type experiments, but our recommendation is to carry one out anyway.
- Starting point: International Principles for Social Impact Assessment by International Association for Impact Assessment (IAIA).

Step 10: Impact assessment

Considerations (cont.):

- IA example dimensions:
  - *Identification of indirect harm:* e.g. what are potential negative impacts of a system of this type being commercially or otherwise deployed?
  - *Equality impact assessment*, e.g.:
    - Are the evaluators representative of the wider population?
    - Does the system type have the potential to create or perpetuate inequalities?

Task: Create a **record of impact assessment** carried out, based on what resources, and with what result.

Example: None that we are aware of.

Step 11: Ethical review

Context:

- All scientific experiments should undergo an ethical review process – principle established in The Nuremberg Code after WW2.
- The required level of review will differ depending on the organisation researchers are affiliated with, and on the type of experiment.
- Most institutions and funding bodies will require an experiment to undergo review by a **research ethics committee** if the experiment involves human participants.
    - Some may have self-review processes for defined types of experiment.
    - The research ethics committee will likely want to see the experiment design.
- Have also encountered institutions without an ethics process!

Considerations:

- Even if your institution does not have a process, you should still self-review.

Step 11: Ethical review

Considerations (cont.):

- In most human evaluations of NLP systems it should be possible to fully mitigate harm, although there are exceptions such as:
  - When the system outputs contain toxic or otherwise upsetting text, e.g., in the evaluation of Mehrabi et al. (2022) which contains a trigger warning.
  - Testing a system that provides feedback about a user's literacy and numeracy skills (Williams & Reiter 2008).

Task:

- Follow own organisation's rules/regulations relating to ethical review of research, typically this requires researchers: (i) to complete a self-assessment checklist, (ii) if indicated by checklist results, to submit the planned research to an ethical review at the indicated level.
- Your funding body may also have ethical guidelines, e.g. UK Research and Innovation (UKRI); Science Foundation of Ireland (SFI).
- If organisation has no ethical review process, carry out ethical self review.
- In both cases create **record of ethical review process and outcome**.

Resources for ethical self-review

- European Commission – Self assessment guidelines, Esp. *Section 2: Humans*:
  - Contains an *Ethics issues checklist* similar to those that an research ethics committee would ask researchers to complete.
  - Also provides details on what researchers should do to address issues, including how to write consent forms and participant information sheets.
- The Economic and Social Research Council (ESRC), a part of UK Research and Innovation (UKRI), has a comprehensive guide for research ethics. In particular, there are sections covering:
  - Consent
  - Core principles
  - Data requirements
  - Internet mediated research
  - Risk and benefit

Legal guidelines and requirements

- Depending on the country where the research is being carried out, there may be legal guidelines or requirements, e.g.:
    - In the UK e.g.: the Equality Act 2010.
    - In the US e.g.: Federal Policy for the Protection of Human Subjects ('Common Rule').
- The institutional research ethics committee will usually include questions covering these issues in their ethical review documents.
- NB: Legal compliance does not mean that research is ethical – hence the two are separate considerations.

Some example ethical review questions

- Will informed consent be obtained from participants?
- Does the research include participants who are in a vulnerable position?
- Is there an existing relationship between the researcher and the participants?
  - Teacher-student
  - Employer-employee (traditional or via crowd platform)
- Are participants excluded based on protected characteristics?
- How will potential participants be recruited?
- Will participants be able to withdraw from the experiment at any time:
  - Can they withdraw their data at any time?
  - Will they still be entitled to any reward for participating?
- Might the research be psychologically harmful to participants, or through repetitiveness impose an unreasonable burden on them?

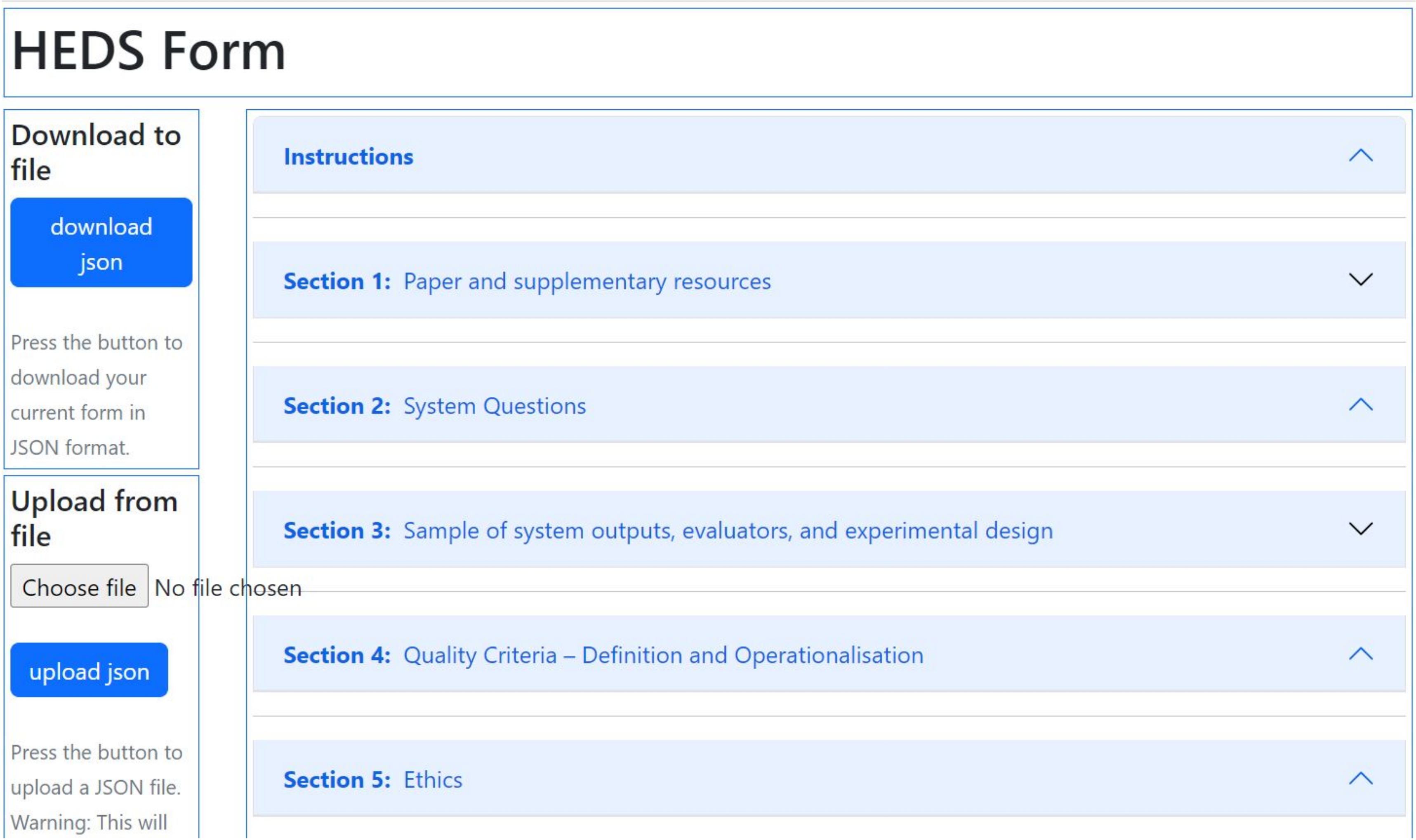
Overview

Unit 4: Experiment Design

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Design decisions and experiment properties
3. Step 3: Output sampling and normalisation
4. Step 4: Rating instrument
5. Step 5: Type and characteristics of evaluators
6. Step 6: Evaluator recruitment and training
7. Step 7: Conditions under which evaluators carry out experiment
8. Step 8: Quality assurance methods
9. Steps 10 and 11: Impact assessment and ethical review
10. Completing human-evaluation datasheet (HEDS)
11. Unit summary and pointers to other units
12. References

Completion of HEDS datasheet

- Once Phase I (Experiment Design) is complete, the experiment is fully specified in its first version.
- It can then be captured by completing a human evaluation datasheet (HEDS).
- Some details may change as a result of piloting and QA testing, requiring updates to the HEDS sheet.
- We have created a tool to make completion and updates easier which we demo next.



HEDS Form

Download to file

Press the button to download your current form in JSON format.

Upload from file

No file chosen

Press the button to upload a JSON file. Warning: This will

Instructions

Section 1: Paper and supplementary resources

Section 2: System Questions

Section 3: Sample of system outputs, evaluators, and experimental design

Section 4: Quality Criteria – Definition and Operationalisation

Section 5: Ethics

Overview

Unit 4: Experiment Design

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Design decisions and experiment properties
3. Step 3: Output sampling and normalisation
4. Step 4: Rating instrument
5. Step 5: Type and characteristics of evaluators
6. Step 6: Evaluator recruitment and training
7. Step 7: Conditions under which evaluators carry out experiment
8. Step 8: Quality assurance methods
9. Steps 10 and 11: Impact assessment and ethical review
10. Completing human-evaluation datasheet (HEDS)
11. Unit summary and pointers to other units
12. References

Unit summary and pointers to other units

- Unit 4 looked at creating the **experiment design** that needs to be combined with a previously selected evaluation measure *m* for a fully specified evaluation method E_m .
- Can think of this as specifying ***how*** we evaluate *m*, whereas Unit 3 dealt with ***what*** is being evaluated, namely the set of evaluation measures *m* .
- We broke experiment design down into different aspects called **experiment (design) properties**, and looked at how they are specified via 11 design steps, and then captured in a HEDS sheet.
- Unit 3 steps specified **quality criteria**, **evaluation modes**, and set of **systems** to evaluate.
- The steps covered in Unit 4 specified:
  - Number of outputs and evaluators and how assigned
  - Output sampling method
  - Rating instrument
  - Type and (qualifying) characteristics of evaluators

Unit summary and pointers to other units

- The steps covered in Unit 4 specified (cont.):
  - Evaluator recruitment, training, instructions.
  - Conditions under which evaluators carry out experiment.
  - Quality assurance methods.
- Unit 5 covers methods for analysis of responses.
- Unit 4 also covered carrying out:
  - A (recommended) impact assessment.
  - The (in most cases obligatory) ethical review.

Overview

Unit 4: Experiment Design

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. Design decisions and experiment properties
3. Step 3: Output sampling and normalisation
4. Step 4: Rating instrument
5. Step 5: Type and characteristics of evaluators
6. Step 6: Evaluator recruitment and training
7. Step 7: Conditions under which evaluators carry out experiment
8. Step 8: Quality assurance methods
9. Steps 10 and 11: Impact assessment and ethical review
10. Completing human-evaluation datasheet (HEDS)
11. Unit summary and pointers to other units
12. References

References

Essential

Disentangling the Properties of Human Evaluation Methods: A Classification System to Support Comparability, Meta-Evaluation and Reproducibility Testing. A Belz, S Mille, D Howcroft. International Natural Language Generation Conference 2020 (INLG'20)

The Human Evaluation Datasheet: A Template for Recording Details of Human Evaluation Experiments in NLP. Anastasia Shimorina and Anya Belz. 2022. 2nd Workshop on Human Evaluation of NLP Systems (HumEval).

QCET: An Interactive Taxonomy of Quality Criteria for Comparable and Repeatable Evaluation of NLP Systems. A Belz, S Mille, Craig Thomson. INLG 2024, to appear.

Further reading

de Leeuw et al. (eds.) (2008) International Handbook of Survey Methodology.

Ruan et al. (2024) Defining and Detecting Vulnerability in Human Evaluation Guidelines: A Preliminary Study Towards Reliable NLG Evaluation.

Generating basic skills reports for low-skilled readers. Sandra Williams and Ehud Reiter. 2008. Natural Language Engineering, 14(4), pp. 495–525.

References

Further reading (cont.)

Robust Conversational Agents against Imperceptible Toxicity Triggers. Ninareh Mehrabi, Ahmad Beirami, Fred Morstatter, and Aram Galstyan. 2022. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics, Seattle, United States

Comparing automatic and human evaluation of NLG systems.
Belz, A., & Reiter, E. (2006, April). In *11th conference of the european chapter of the association for computational linguistics* (pp. 313-320).

The 2023 WebNLG shared task on low resource languages overview and evaluation results (WebNLG 2023).
Cripwell, Liam, et al. *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*. 2023.

Common Flaws in Running Human Evaluation Experiments in NLP.
Thomson, C., Reiter, E., & Belz, A. (2024). *Computational Linguistics*, 1-11.

The TUNA-REG Challenge 2009: Overview and evaluation results.
Gatt, A., Belz, A., & Kow, E. (2009). Association for Computational Linguistics.

References

Further reading (cont.)

Findings of the 2019 conference on machine translation (WMT19).
Barrault, Loïc, et al. ACL, 2019.

Generating basic skills reports for low-skilled readers.
Williams, S., & Reiter, E. (2008). *Natural Language Engineering*, *14*(4), 495-525.