

Human Evaluation of NLP System Quality

INLG Tutorial, 24th September 2024

Unit 7: Experiment Execution

[Link to Unit 7 Resources](#)

Overview

Unit 7: Experiment Execution

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. From implementation to execution
3. Pre-final execution
4. Preregistration and final execution
5. Ethics and fair treatment of participants
6. Post-final execution for reproducibility testing
7. Unit summary and pointers to other units
8. References

Overview

Unit 7: Experiment Execution

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. From implementation to execution
3. Pre-final execution
4. Preregistration and final execution
5. Ethics and fair treatment of participants
6. Post-final execution for reproducibility testing
7. Unit summary and pointers to other units
8. References

Unit aims, learning outcomes, prerequisites from other units

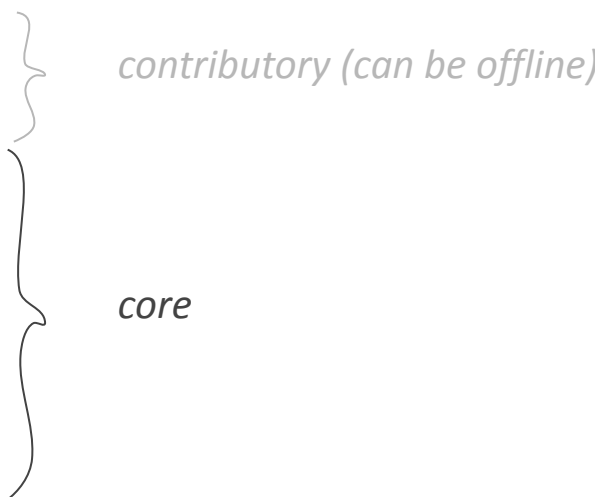
- The aims of Unit 7 are:
 - To look at the process of executing an implemented experiment design.
 - To discuss testing and piloting.
 - To introduce reproducibility of results.
- After completion of the unit, participants will be able to:
 - Understand the steps involved in executing and implemented experiment design, both in pre-final and final execution.
 - Be aware of ethical issues that may arise during experiment execution.
- Prerequisites:
 - All previous units, but especially Unit 5 (Analysis) and Unit 7 (Implementation).

Overview

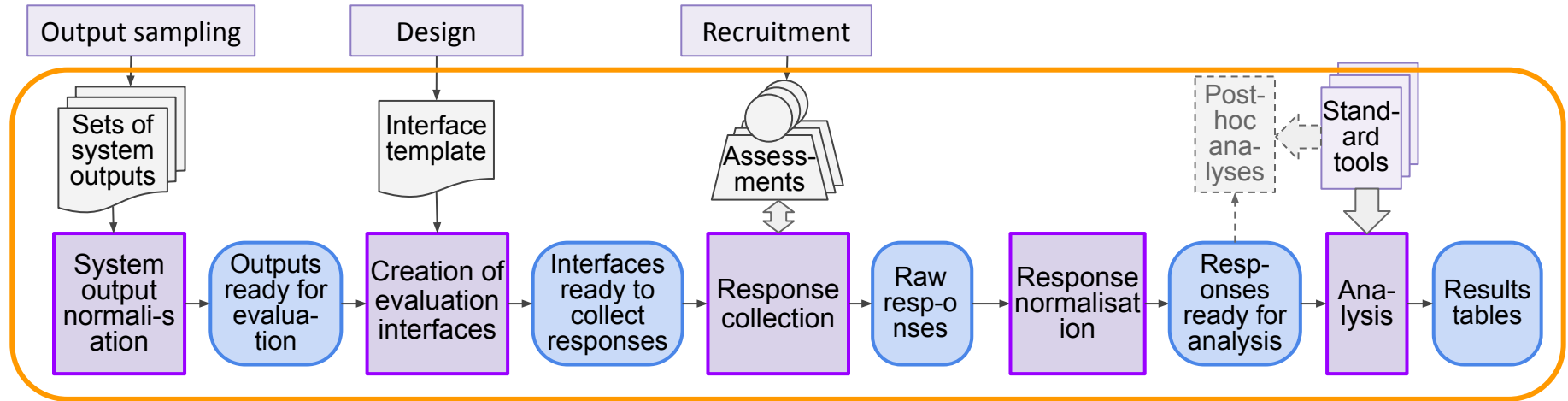
Unit 7: Experiment Execution

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. From implementation to execution
3. Pre-final execution
4. Preregistration and final execution
5. Ethics and fair treatment of participants
6. Post-final execution for reproducibility testing
7. Unit summary and pointers to other units
8. References

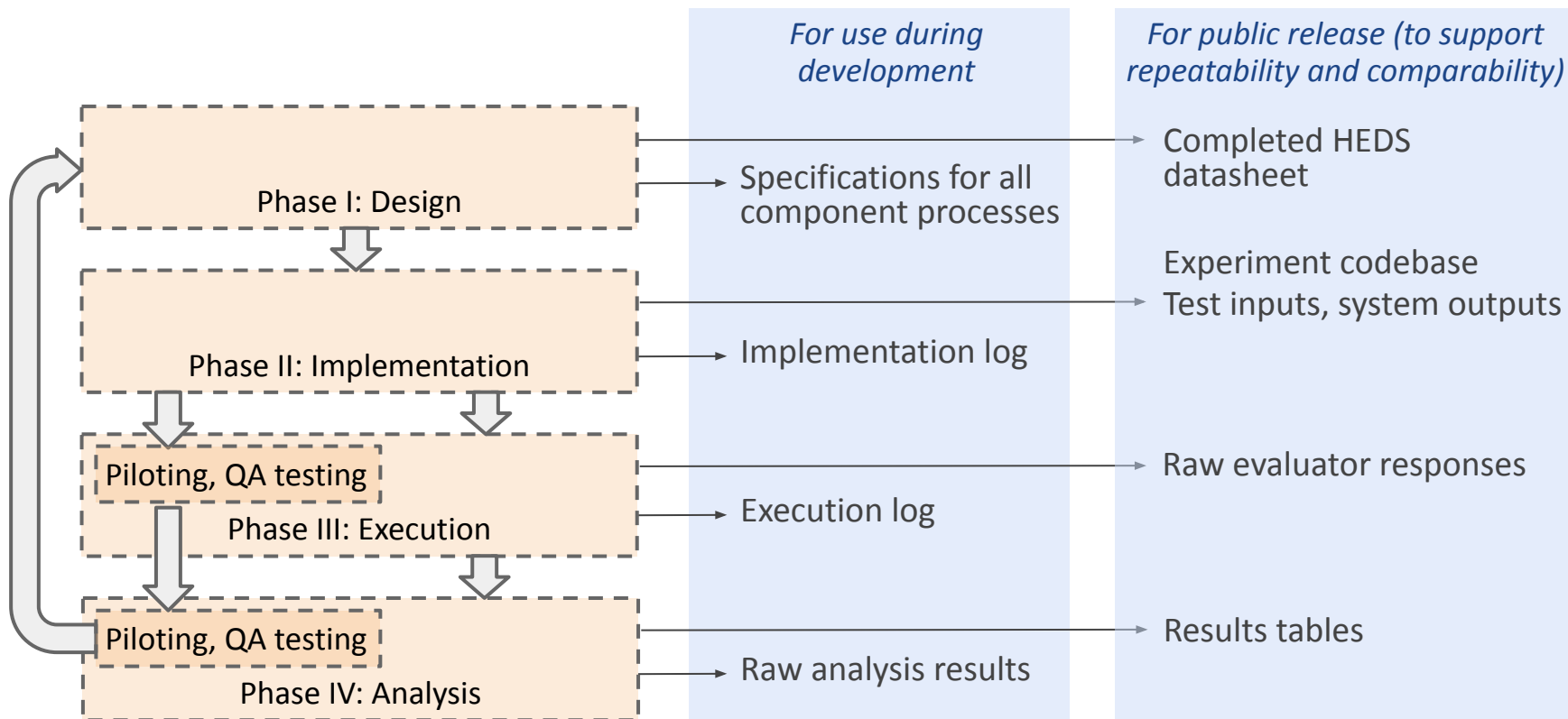
From implementation to execution

- We will assume that output sampling and recruitment have been completed offline and focus on the five core processes:
 - Output sampling
 - Recruitment
 - Output normalisation
 - Evaluation interface creation
 - Response collection
 - Response normalisation
 - Analysis
- 
- The diagram uses curly braces to group the processes. A small brace groups 'Output sampling' and 'Recruitment', with the label *contributory (can be offline)* to its right. A larger brace groups 'Output normalisation', 'Evaluation interface creation', 'Response collection', 'Response normalisation', and 'Analysis', with the label *core* to its right.
- If full pipelining was found to be feasible in Phase II, the five processes can all be executed at once, from evaluation items to results tables.

Component process diagram for reference



Phase diagram for reference



Execution topics

- Experiment design (Phase I, Units 3 and 4) has been fully implemented (Phase II, Unit 6) in advance.
- Therefore, the execution of the experiment itself for the purpose of collecting responses should be straightforward: written procedures are followed and code pipelines are run.
- However, there are still some additional issues that we should be aware of when executing our experiments.
- We look at three different purposes for which experiments are executed:
 - Testing (**pre-final execution**);
 - Running the actual experiment as it will be reported (**final execution**); and
 - Reproducibility testing (**post-final execution**).

Execution topics: Pre-final execution

- When discussing **pre-final execution**, we look at:
 - Interface testing,
 - Pilot experiments (typically with smaller numbers of participants and evaluation items),
 - Checks for intra and inter-annotator agreement,
 - Obtaining qualitative feedback from pilot participants, and
 - Updating the experiment design (by going back to previous phases if necessary) based on issues raised during pre-final execution.

Execution topics: Final execution

- When discussing **final execution**, we look at:
 - Creating the **pre-registration**, beyond which point the design and implementation become fixed.
 - Keeping an **experiment log** which can serve as a record of the experiment steps being executed. If there were unavoidable changes that had to be made during final execution they should also be recorded here, although these may be serious, requiring an updated pre registration and new experiment.
 - Issues that come up during experiment execution, e.g. how participants should be treated in an ethical and fair way.

Execution topics: Post-final execution

- When discussing **post-final execution**, we look at:
 - Reproducibility testing: a special case of execution where another team executes the experiment, ideally having access to all resources shown in the box on the right of the phase diagram, with the aim of comparing similarity of results.
 - Terminology used for reproducibility studies (Belz, 2022).
 - How the degree of reproducibility between two or more studies can be measured, using extended Quantified Reproducibility Assessment (QRA++) (Belz et al., 2022; Belz and Thomson, 2023, 2024).

For public release (to support repeatability and comparability)

Completed HEDS
datasheet

Experiment codebase
Test inputs, system outputs

Raw evaluator responses

Results tables

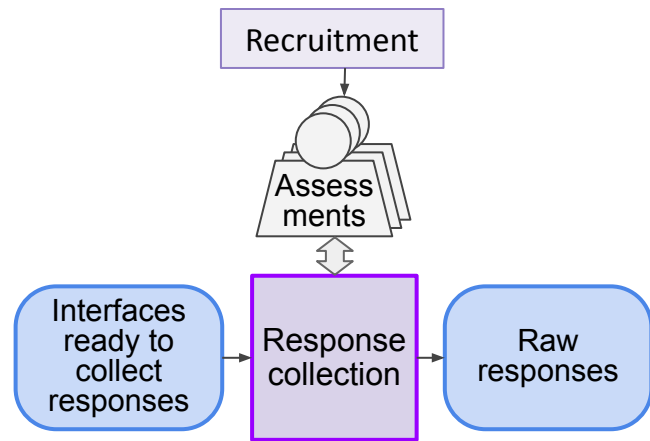
Overview

Unit 7: Experiment Execution

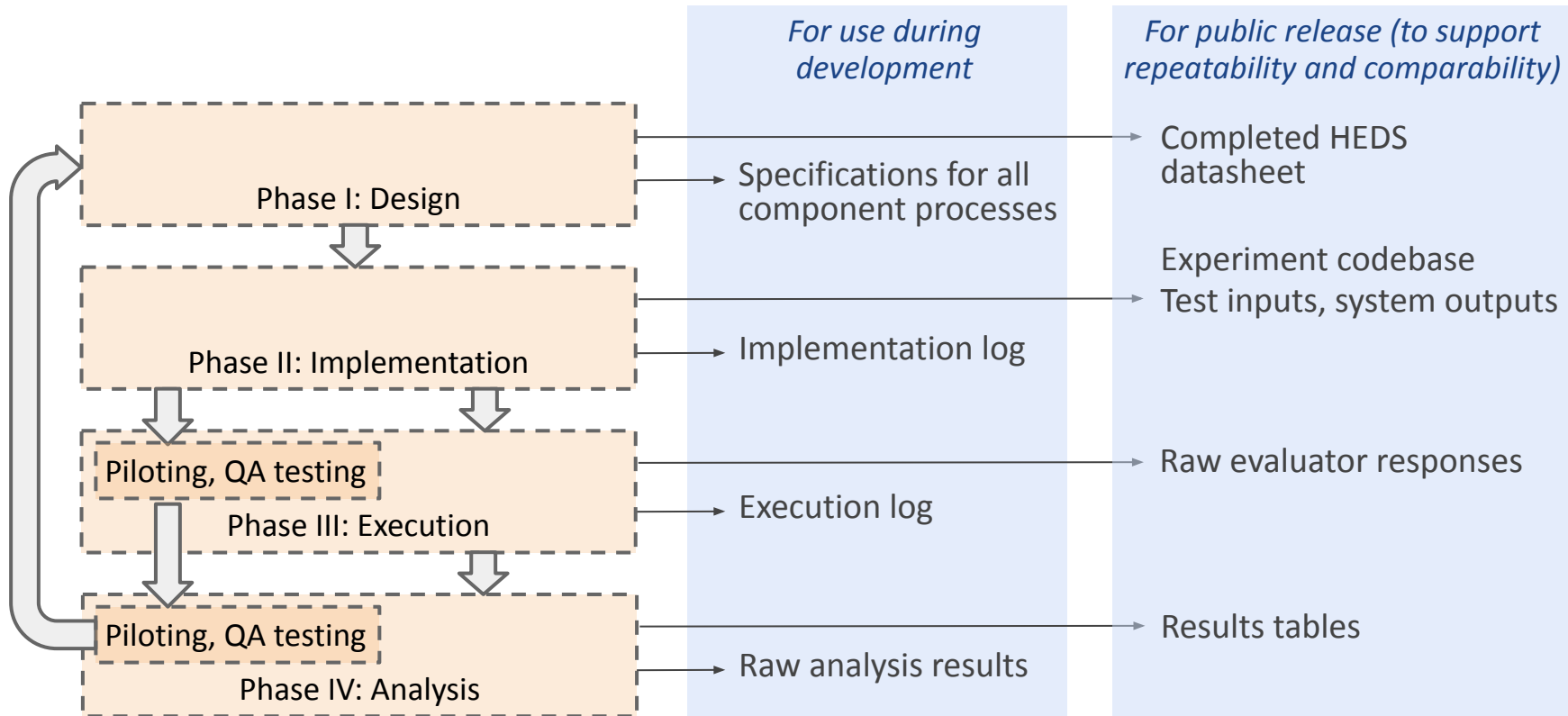
1. Unit aims, learning outcomes, contents and prerequisites from other units
2. From implementation to execution
3. Pre-final execution
4. Preregistration and final execution
5. Ethics and fair treatment of participants
6. Post-final execution for reproducibility testing
7. Unit summary and pointers to other units
8. References

Pre-final execution

- Everything before **Interfaces ready to collect responses**, and after **Raw responses**, should have been implemented in a code pipeline, which can simply be run.
- For pre-final execution, and for (final) execution, we will focus on the response collection component process.



Steps in creating and running a human evaluation



Pre-final execution

- Iterate as necessary:
 - Perform interface testing.
 - Run response collection component process with smaller number of evaluators and evaluation items as pilot experiment.
 - Test pilot responses for inter and intra-annotator agreement.
 - Collect feedback from pilot evaluators re understandability and complexity of task.
 - If feasible, test for reproducibility.
 - Collate and implement improvements.

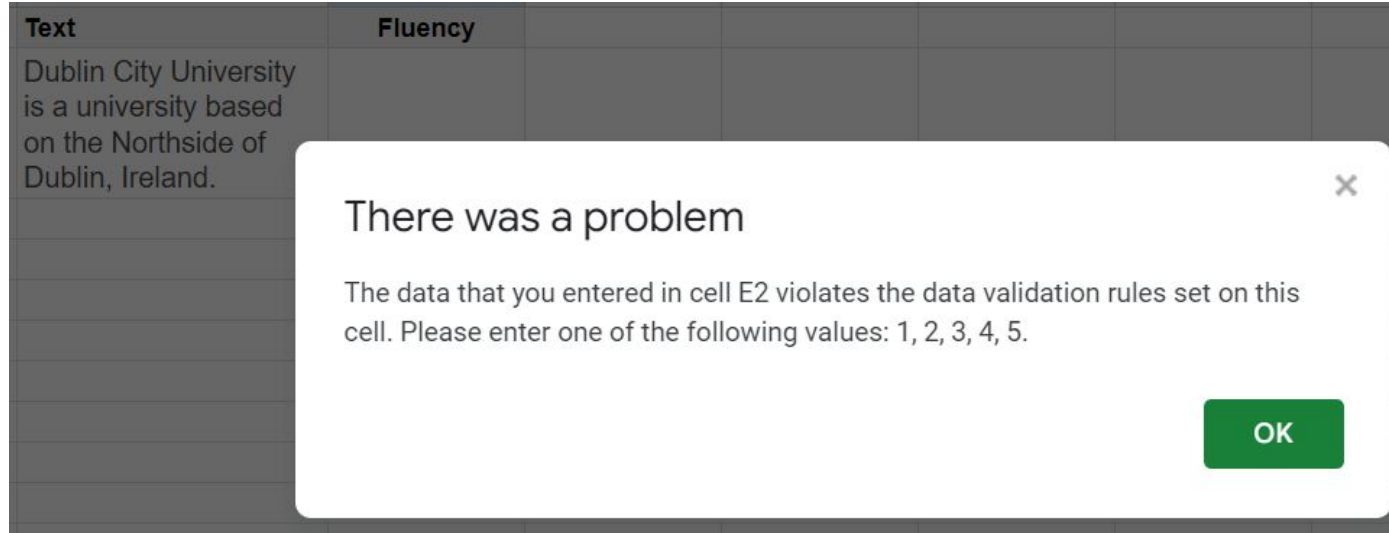
Interface testing – data validation

Examples of tests that can be performed.

- Create a list of valid and invalid responses, and check whether the interface accepts them.
- Check that input fields accept long strings, or limit the input text to the size of the field.
- Systematically check any links or buttons to ensure they work.
- Check that the user cannot submit a form that has errors, or, in the case of spreadsheets, ensure that there is a clear messages shown when the form is incomplete.

D	E
Please enter a score for each text	
Text	Fluency
Dublin City University is a university based on the Northside of Dublin, Ireland.	5 ▼
NYU Stern is the business school of New York University	▼

Interface testing – structure and automation



- Ensure that the structure of the interface is immutable, i.e., protect a spreadsheet such that participants cannot change its structure and can only enter content in the response fields.
- For automated testing, use tools such as [Selenium](#) to simulate user activity.

Pilot experiment

- Exactly like the planned experiment, but on a smaller scale:
 - Fewer evaluation items (using a sample that is distinct from the main experiment).
 - Fewer participants but with the same characteristics.
 - It may be possible to use participants with slightly different characteristics, e.g., for an initial pilot within your lab. However, this may make any inter and intra-annotator agreement scores less reliable predictors of agreement in the main experiment.
- Execute the pipeline(s) of component processes.

Pilot inter and intra-annotator agreement

- We would like this to be as high as possible, and piloting/revising the experiment is a good way to attain this. Some guideline values are:
 - E.g., Krippendorff's alpha ≥ 0.667 .
 - Krippendorff (2004) suggests a minimum of 0.667 where tentative conclusions are acceptable, otherwise 0.800.
 - Thresholds for other metrics exist:
 - Landis & Koch, (1977) suggest ≥ 0.600 for substantial agreement with Fleiss Kappa.
- Calculate intra and inter-annotator agreement as specified in the design, using code from the implementation.
- Please see Unit 5 for definitions of agreement metrics.

Collect feedback from pilot evaluators

Some example questions they could be asked:

- Did they understand the task that was asked of them?
- Were any parts of the instructions ambiguous or otherwise unclear?
- Did they have any problems using the interface?
- Did they encounter any other issues?

Collate and implement improvements

- Based on feedback from participants of the pilot, alter the design and implementation of the experiment.
- If further revision to the design is required:
 - Adjust the design/implementation then run another pilot.
 - Remember that not doing the experiment is an option.
- If the pilot was successful:
 - Changes can be made without running another pilot, although these should be minimal, such as correcting typos and minor bugs.

Overview

Unit 7: Experiment Execution

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. From implementation to execution
3. Pre-final execution
4. Preregistration and final execution
5. Ethics and fair treatment of participants
6. Post-final execution for reproducibility testing
7. Unit summary and pointers to other units
8. References

Final execution

- Pre-register the experiment.
- All component processes are tested and clearly defined by this stage.
- Run response collection component process with the final number of participants and evaluation items.
- Keep an experiment log.

Preregistration

- The preregistration should be created once the design is finalised and the implementation is complete, tested, and the experiment has been piloted.
- Sites such as aspredicted.org for a simple registration, or <https://help.osf.io/article/145-preregistration> for a more comprehensive one, can be used.
- Uploading the design, along with all code and documentation, to [GitHub](https://github.com) is good practice. The repository can be kept private initially but will be time stamped and can be referenced in aspredicted.org.

The experiment log

- It is good practice to keep an experiment log.
- The log might take the form of a spreadsheet with the following columns:
 - **Phase:** The current phase of the experiment.
 - **Component process:** The current component process within the phase.
 - **Completion Date:** The date the component process was completed.
 - **Completion Time:** The time the component process was completed.
 - **Changes:** Any changes to the experiment that had to be made.
 - **Other notes:** Any other notes that might help you or another researcher in the future.

Changes to the experiment

- In order to complete the experiment it might be necessary to change the experiment implementation. E.g.,
 - Despite testing, a results script failed to run.
 - It proved impossible to recruit participants of the required characteristics and the most suitable alternative was found. For example, the experiment required 10 post docs, but 9 post-docs and one final year PhD student were recruited.
- All such changes should be recorded and noted in any research papers.
- Full details should be included as an appendix or supplementary material.

Overview

Unit 7: Experiment Execution

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. From implementation to execution
3. Pre-final execution
4. Preregistration and final execution
5. Ethics and fair treatment of participants
6. Post-final execution for reproducibility testing
7. Unit summary and pointers to other units
8. References

Ethics and fair treatment of participants.

- Follow the experiment design and any feedback from the ethical review (see Unit 4).

Payment of participants

- It is reasonable to compensate participants, either via direct payment or by gift voucher, for all but the shortest (10-15 minute) experiments.
- Payment at or above the minimum wage in the country where the research is being carried out is a good practice.
- If recruiting participants from another country, consider using the minimum wage there if it is higher.

Treatment of crowd workers

- All participants must be treated fairly and with a high standard of ethics from the outset.
- Crowd workers *are* participants (Shmueli et al. 2021).
- Fort et al. (2011) and Shmueli et al. (2021), among others, have expressed concerns about the ethical treatment of crowd-sourced participants on platforms such as Amazon Mechanical Turk. Shmueli et al. (2021) report that:
 - Some students are taught that experiments with crowd-sourced participants do not require an ethical review, *they do*.
 - Some ethical review boards have been slow to adapt to the new types of issue raised by crowd platforms.

Treatment of crowd workers

- Some participants may be under pressure to find additional income, or have no other opportunities due to a lack of local job opportunities, making this job a lifeline.¹
- There is a power imbalance between participants on platforms such as Mechanical Turk, where requesters can reject participants work, not paying them, with little oversight. This is not OK.
- Some platforms, such as Prolific, make an effort to address this.
- It's best to pay participants for work already completed even if they fail an attention or other check. Bad faith is difficult to establish beyond reasonable doubt. It may not be their fault if they cannot complete your experiment.
 - Platforms such as Prolific have attention check policies.

¹“AI: Ghost workers demand to be seen and heard”, March 2021: <https://www.bbc.co.uk/news/technology-56414491>

Overview

Unit 7: Experiment Execution

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. From implementation to execution
3. Pre-final execution
4. Preregistration and final execution
5. Ethics and fair treatment of participants
6. Post-final execution for reproducibility testing
7. Unit summary and pointers to other units
8. References

Reproducibility

- Experimental results should be reproducible.
- ACM: “An experimental result is not fully established unless it can be independently reproduced.”¹
- If they’re not (i.e. if get different results when you run the same experiment again) then conclusions based on them are unsafe.
- At best 1 in 6 papers provide sufficient resources to attempt a repeat of a study (Belz et al., 2023)
- Using the approach, and in particular sharing resources, as recommended in this tutorial, will ensure your experiments are repeatable.

From 4-phase diagram:

For public release (to support repeatability and comparability)

Completed HEDS
datasheet

Experiment codebase
Test inputs, system outputs

Raw evaluator responses

Results tables

1. <https://www.acm.org/publications/policies/artifact-review-and-badging-current>

Standard scientific definitions

- Taken from the International Vocabulary of Metrology:
 - **repeatability**: the precision of measurements of the same or similar object obtained under the same conditions.
 - **reproducibility**: the precision of measurements of the same or similar object obtained under different conditions.
- Term 'reproduction study' refers to any study that tests repeatability or reproducibility in the above sense.

Quantified Reproducibility Assessment (QRA++)

- Different types of results reported in papers for evaluation experiments:
 - (a) Type I results: single numerical scores, e.g. mean quality rating, error count, etc.
 - (b) Type II results: sets of numerical scores, e.g. set of Type I results.
 - (c) Type III results: categorical labels attached to text spans of any length.
 - (d) Qualitative conclusions/findings stated explicitly in the original paper.
- How QRA++ computes quantified reproducibility assessments for each type of result:
 - (a) Type I results: CV^* (coefficient of variation debiased for small samples).
 - (b) Type II results: Pearson's r , Spearman's ρ
 - (c) Type III results: Multi-rater: Fleiss's κ ; Multi-rater, multi-label: Krippendorff's α .
 - (d) Conclusions/findings: Proportion of pairwise ranks that are / are not confirmed in the reproduction experiment.

Example Type I

Absolute evaluation of fluency on a scale of 1 to 5.

System	Orig	Repro 1	Repro 2	
SVM	3.71	3.12	3.02	$CV^* = 19.96$
GeDi	3.20	2.57	2.40	$CV^* = 29.90$
DExpert	2.33	2.28	1.81	$CV^* = 26.87$

Lower CV^* is better. A score in the range of 20-30 is medium-poor (the per-study scores differ).

Example Type II

System	Orig	Repro 1	Repro 2
SVM	3.71	3.12	3.02
GeDi	3.20	2.57	2.40
DExpert	2.33	2.28	1.81

$$r = 0.95$$
$$\rho = 1.00$$

Strong correlations, despite the difference in per-study system scores.

Example Type II

System	Orig	Repro 1	Repro 2
SVM	3.71	3.12	3.02
GeDi	3.20	2.57	2.40
DExpert	2.33	2.28	1.81

$$r = 0.99$$
$$\rho = 1.00$$

Strong correlations, despite the difference in per-study system scores.

Example Type II

System	Orig	Repro 1	Repro 2
SVM	3.71	3.12	3.02
GeDi	3.20	2.57	2.40
DExpert	2.33	2.28	1.81

$$r = 0.99$$
$$\rho = 1.00$$

Strong correlations, despite the difference in per-study system scores.

Example Type III

- The lack of available of raw responses from original experiments is usually a barrier in carrying these out.
- Calculated like inter-annotator agreement, but inter-study.

Example Type IV

A => best system, **B** => second, **C** => worst.

System	Orig	Repro 1	Repro 2
SVM	3.71 A	3.12 A	3.02
GeDi	3.20 B	2.57 B	2.40
DExpert	2.33 C	2.28 C	1.81

In this case, the order is the same in all studies, so the pairwise (between studies) rank is confirmed for 3 of 3 ranks.

Example Type IV

A => best system, **B** => second, **C** => worst.

System	Orig	Repro 1	Repro 2
SVM	3.71 A	3.12	3.02 A
GeDi	3.20 B	2.57	2.40 B
DExpert	2.33 C	2.28	1.81 C

In this case, the order is the same in all studies, so the pairwise (between studies) rank is confirmed for 3 of 3 ranks.

Example Type IV

A => best system, **B** => second, **C** => worst.

System	Orig	Repro 1	Repro 2
SVM	3.71	3.12 A	3.02 A
GeDi	3.20	2.57 B	2.40 B
DExpert	2.33	2.28 C	1.81 C

In this case, the order is the same in all studies, so the pairwise (between studies) rank is confirmed for 3 of 3 ranks.

Overview

Unit 7: Experiment Execution

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. From implementation to execution
3. Pre-final execution
4. Preregistration and final execution
5. Ethics and fair treatment of participants
6. Post-final execution for reproducibility testing
7. Unit summary and pointers to other units
8. References

Unit summary and pointers to other units

- Unit 7 looked at the **experiment execution**, based on the **experiment implementation** from Unit 6 that was in turn based on the **experiment design** from Unit 4.
- Pre registration, and the need for it was discussed.
- We looked at different ways in which the experiment implementation could be executed:
 - Testing (**pre-final execution**);
 - Running the actual experiment as it will be reported (**final execution**); and
 - Reproducibility testing (**post-final execution**).

Unit summary and pointers to other units (cont.)

- The **experiment design** from Unit 4 should have been made to a high ethical standard, there was a discussion in this unit of ethical issues that might arise at the time of **experiment execution**.
- Standard definitions for reproducibility testing were given, as well as methods for Quantified Reproducibility Assessment (QRA++)

Overview

Unit 7: Experiment Execution

1. Unit aims, learning outcomes, contents and prerequisites from other units
2. From implementation to execution
3. Pre-final execution
4. Preregistration and final execution
5. Ethics and fair treatment of participants
6. Post-final execution for reproducibility testing
7. Unit summary and pointers to other units
8. References

References

Essential:

[A Metrological Perspective on Reproducibility in NLP.](#)

A Belz. *Computational Linguistics* 2022; 48 (4): 1125–1135.

[The 2024 ReprONLP Shared Task on Reproducibility of Evaluations in NLP: Overview and Results.](#)

Belz & Thomson, HumEval-WS 2024.

[Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP.](#)

A Belz, C Thomson, E Reiter, S Mille. Findings of ACL 2023, 3676-3687, 2023.

[Beyond fair pay: Ethical implications of NLP crowdsourcing.](#)

Shmueli, Boaz, et al. *arXiv preprint arXiv:2104.10097* (2021).

References

Further reading:

[Generating scientific definitions with controllable complexity.](#)

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317.

[Report for ReproHum project 0033: Comparable Relative Results with Lower Absolute Values in a Reproduction Study.](#)

Yiru Li, Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2024. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*. Association for Computational Linguistics.

[ReproHum:# 0033-03: How Reproducible Are Fluency Ratings of Generated Text? A Reproduction of August et al. 2022.](#)

Emiel van Miltenburg, Anouck Braggaar, Nadine Braun, Martijn Goudbeek, Emiel Krahmer, Chris van der Lee, Steffen Pauws, and Frédéric Tomas. 2024. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*. Association for Computational Linguistics.

References

Further reading (cont.):

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. ELRA and ICCL.

[Measuring the reliability of qualitative text analysis data.](#)

Krippendorff, K. (2004). *Quality and quantity*, 38, 787-800.

[An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers.](#)

Landis, J. R., & Koch, G. G. (1977). *Biometrics*, 363-374.