The Journal of Open Source Software

Comparative Judgement Interface: A Data Collection Interface for Comparative Judgement Studies

Catherine Smith[1¶], **Fabián Hernández**[2], **Bertrand Perrat**[3], **Adrian Dante Garcia**[1], **and Rowland G. Seymour**[1]

1 University of Birmingham, United Kingdom **2** Independent Software Engineer, Costa Rica **3** Independent Software Engineer, United Kingdom ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 11 November 2025
Published: unpublished

Summary

Comparative judgement is a research method where participants (referred to as judges) are asked to make pairwise judgements of items based on a given criteria, for example which of two areas has the higher rate of deprivation, or which of two student assignments shows a better understanding of a particular topic. These pairwise judgements can then be used to rank all the items in the study based on the criteria. Pairwise judgements have been shown to be easier for people to make than alternatives such as making an absolute judgement based on a scale or putting a whole series of items in order based on a criteria (Jones & Davies, 2023). These models are gaining popularity in social sciences but the options for data collection remain limited. The comparative-judgement-interface is a Flask app that provides a specialist interface for comparative judgement which is domain agnostic and highly configurable in terms of its behaviour and the text presented to the judge. This means that the same code can be configured to collect comparative judgement data on different topics using different languages making it applicable to a wide range of research projects. The website complies with WCAG 2.2 level AA and works on smaller screens such as phones and tablets as well as larger screens. It can be configured without any programming knowledge using JSON, or a combination of JSON and a CSV file. Any set of images can be used in the app. Studies can either be configured on the command line or, if enabled, via a web based admin interface. All of the data collected is stored in a SQLite3 database and an export to CSV files is provided. Example configurations are provided as a starting point for users. The interface has been used to collect data for a number of studies, for example Seymour et al. (2022); Seymour & Hernandez (2024); Seymour et al. (2025) and has now been deployed in partnership with seven councils and local authorities in the UK.

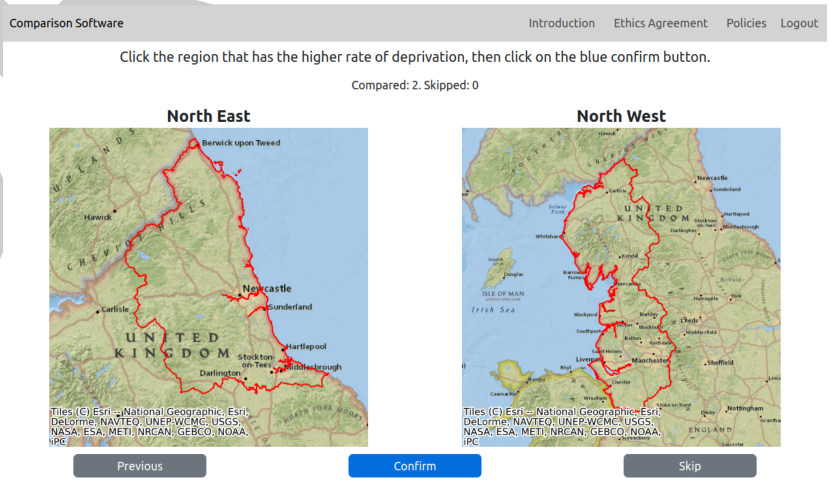
Statement of Need

Mainstream survey collection applications, such as KoBo Toolbox or Online Surveys, do not currently allow for comparative judgement questions to be included in surveys. This limits the application of comparative judgement in the social sciences. The open source python package PsychoPy (Peirce et al., 2019) can be used to run comparative judgement studies. This software is not specific to comparative judgement and allows a lot of different types of studies to be created. While this is an advantage for research that needs to combine different techniques, it makes setting up a comparative judgement study more complex. The PsychoPy package is also aimed at academics and researchers in experimental psychology. In the education assessment domain several software platforms exists, but they are highly tailored to this specific application and are all proprietary. For example, NoMoreMarking (Jones & Wheadon, 2015) provides, proprietary but free to use software that is designed for schools and other education providers to run studies about student assessment. While custom studies can be created in the system, the design of these is tailored to the education context and is not applicable to all research questions. For example, to track user engagement, each judge needs to be sent the url and register with their email address. This makes it unsuitable for open or anonymous studies. We required an app that was flexible in terms of the items that could be compared and the wording for the research question. Our comparative-judgement-interface aims to make comparative judgement studies simpler, more efficient and more attractive to run across a wide range of disciplines.

Current comparative judgement data collection software implements standard experimental designs, limiting the way in which surveys can be run. Our app allows for three types of experimental design configuration. The first is to split the study items into multiple groups and ask the judges which groups they are familiar with. This makes the data collection process more efficient than asking judges to compare objects they are not familiar with, which can increase survey fatigue. The second is allowing the probability that different items are compared against each other to be customised. The default option is to chose objects to compare against each other uniformly at random from the list of all possible pairs, but we also allow probabilities to be specified for each pair, allowing some pairs to be featured more frequently than others. The third is that we allow for comparisons to be tied. This can be useful in studies where there are several items that are difficult to distinguish between.

Configuration Summary

The full list configuration options are explained in our documentation. The app includes the option to configure all of the text that the judge sees on the webpage which means that the system can be configured to work with any language, defaults are provided in English for all text strings. Some of the behaviour of the website can also be tailored to the requirements and experimental design of the study. This includes the ethics and legal information shown the the judges, as well as the experimental design features outlined in the Statement of Need. There is an optional admin interface which allows logged in users to setup and monitor studies via a web interface. An optional API also provides access to selected tables from the database for live analysis.



Click the region that has the higher rate of deprivation, then click on the blue confirm button.

Compared: 2. Skipped: 0

Figure 1: A screenshot of the ranking page from the comparative judgement interface

Acknowledgements and Funding

This work was supported by a UKRI Future Leaders Fellowship [MR/X034992/1].

References

Jones, I., & Davies, B. (2023). Comparative judgement in education research. *International Journal of Research & Method in Education*, *47*(2), 170–181. https://doi.org/10.1080/1743727x.2023.2242273

Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation*, *47*, 93–101. https://doi.org/https://doi.org/10.1016/j.stueduc.2015.09.004

Peirce, J., Gray, J., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). *PsychoPy2: Experiments in behavior made easy*. https://doi.org/10.3758/s13428-018-01193-y

Seymour, R. G., & Hernandez, F. (2024). Scalable bayesian inference for bradley–terry models with ties: An application to honour based abuse. *Journal of Applied Statistics*, 1–18. https://doi.org/10.1080/02664763.2024.2436608

Seymour, R. G., Nyarko-Agyei, A., McCabe, H., Severn, K., Sirl, D., Kypraios, T., & Taylor, A. (2025). Comparative judgement modelling to map forced marriage at local levels. *The Annals of Applied Statistics*, *19*(1), 419–439. https://doi.org/10.1214/24-AOAS1966

Seymour, R. G., Sirl, D., Preston, S. P., Dryden, I. L., Ellis, M. J. A., Perrat, B., & Goulding, J. (2022). The bayesian spatial bradley–terry model: Urban deprivation modelling in tanzania. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *71*(2), 288–308. https://doi.org/10.1111/rssc.12532