LOW STAKES, HIGH REWARD: DEVELOPMENT AND EVALUATION OF AN ONLINE COMPARATIVE JUDGEMENT MARKING SYSTEM

T. Gleave, S. Goodman, S. Canning, D. Prescott

University of Liverpool (UNITED KINGDOM)

Abstract

In the ever-evolving landscape of education, the quest to find more effective, reliable, and efficient methods of assessing students' knowledge and providing constructive feedback has been ceaseless. Over 70,000 articles and 6,000 books have been published on this subject in the last decade alone. In addition, there has been much debate regarding grade reliability for at least the last 100 years [1]. In response, a transformative approach known as *Comparative Judgement* (CJ) is revolutionising the way educators mark assignments and deliver feedback, offering a dynamic alternative to conventional grading practices [2],[3]. Rooted in educational theory and empowered by technological advancements, CJ reimagines the assessment process by embracing the principles of comparison, fairness, and objectivity. The following narrative summarises a formal research project on CJ, which has University research ethical approval.

The primary research aim is to develop, integrate, and implement a peer learning, assessment, and feedback tool, allowing assessment judgements to be made in a 'low stakes' and efficient approach on undergraduate modules delivered at the University of Liverpool, UK. The first stage will begin with developing and integrating the *'CJ Application'* into CANVAS, the university's virtual learning environment (VLE). The application will be used as a learning, assessment, and feedback tool after integration. Staff and students from a range of modules in the School of Life Sciences (SoLS) will have the opportunity to be involved in the project.

Both students and staff will use the pilot version of the 'CJ application', which will generate a vast collection of feedback and a 'rank' of appraised poster submissions. These posters will also be marked independently utilising the 'traditional' approach, and comparisons/analyses will be made between the 'traditional', CJ student, and CJ staff ranks.

Staff and student focus groups will also be facilitated to evaluate the appraisal and marking processes. Full and informed participant consent will be sought for all study elements in alignment with university policy. The research team will also analyse anonymised data from the EVAsys module evaluation data, access analytics data from CANVAS (VLE), and compare the assessment component marks. These records are collected as part of the university's standard module assessment and evaluation procedures.

All data from CANVAS, the focus groups and student module evaluations' EVAsys' will be used to evaluate the CJ application for further development, promoting discussion about potential future CJ utility and dissemination. The risks, challenges and enablers experienced during the pilot research phase will be discussed, as will the potential for wider 'roll out' across the university and HE sectors.

Keywords: Comparative judgement, grade reliability.

1 INTRODUCTION

The School of Life Sciences (SoLS) at the University of Liverpool, United Kingdom, offers world-class, research-led education across the entire life science spectrum, including biomedical sciences, whole organismal biology, molecular genetics, infection and microbiology, structural biology, biological chemistry, ecology, and evolution. The school offers a complete portfolio of undergraduate and postgraduate programmes and delivers world-class education to over 1500 students.

A key ambition for the SoLS is to provide a learning environment that affords opportunities for students to become more confident while developing their independence and critical thinking skills. Academics across the school are mindful of the need to engage students as active partners in their learning journeys, not passive participants, or passengers. Authentic assessments provide opportunities for learners to apply their knowledge and demonstrate their understanding of concepts and principles in 'real' situations and to solve problems seen in the 'real world'. The following narrative will describe 'Comparative Judgment', a pedagogical strategy purported to promote learners' skills while providing a valid and reliable assessment method.

Comparative judgment (CJ) has the potential to provide reliable assessment, enhancing educational practice through the provision of a mechanism which supports formative feedback alongside the opportunity to provide evidence of learning. Unlike traditional criterion-based assessments, the CJ process is grounded on comparisons between two pieces of work viewed contemporaneously. Cohorts of assessors are individually presented with a pair of student works, from which they must select the better of the two. When undertaking this task, the assessor is moved from a position of deciding the mark to be allocated for an individual piece of work in relation to grade-related criteria to one whereby they simply decide which piece of work from the pair in front of them, more strongly evidences learning based on their professional judgment [4]. After individual assessors make these judgments, the combined results are placed in rank order, validating the CJ process for the cohort of assessors [5]. At this point, it is important to note that the highest-ranking piece of work may not necessarily be of high quality, and likewise, the lowest-ranking piece of work may not be of poor quality. Consequently, comparisons undertaken for a particular assessment task across a specific cohort of students can also provide a basis for critical discussion from multiple assessors, giving students access to extended formative or summative feedback from multiple sources. Likewise, students can also use this process by asking them to review pairs of work, determine the 'better' piece of work and provide feedback. Therefore, the CJ tool is utilised as a pedagogical strategy to promote learners' analytical and evaluative skills.

The transfer of the rankings to percentages or grades is achieved after the CJ process has been undertaken, and this can be delivered through various methods, dependent upon the aims of the assessment and the associated learning outcomes [6]. An obvious benefit of the CJ process is that comparisons made by multiple assessors rule our personal preferences, leading to more consistent judgment across assessors [7]. In contrast to a more traditional piece of assessment, marking a single assessment is undertaken by one or two markers, providing limited academic judgement of the piece. Issues arising from this are the variability of the marking criteria and the feedback provided to the learners, which is in one individual's opinion. This pertains to the issue of assessment reliability, marking and feedback, and there has been much debate on grade reliability for the last 100 years [1].

Over recent years, some institutions have incorporated digital tools to support their use of comparative judgment. It is well established in the literature that there is a growing demand across educational contexts for using digital tools in assessment [6]. However, most digital assessment tools rely on more traditional means of assessing student achievement instead of adding value to the student experience or attempting to reduce the educator workload. Likewise, the importance of grades [8] and learner engagement in feedback [9] have been well documented in the literature, and therefore, it is imperative that both learners and educators are fully engaged in the implementation of any assessment innovation.

We believe that Comparative Judgment has the potential to stand as a 'beacon' of innovation in the realm of education by embracing the principles of comparison, fairness, and objectivity. This work aims to provide educators with a transformative tool to assess and effectively guide their students' learning journey. This pilot project focuses on developing a digital CJ tool and its application through a low-stakes project where academics, students and digital experts work together.

2 METHODOLOGY

Research Aim: To develop, integrate and implement a bespoke Comparative Judgment assessment tool.

This pilot project sought to gather the thoughts and opinions of the end users (students and staff) and the development team, as all parties worked together on the tool's development, implementation, and subsequent refinement. Ethical approval was sought and gained from the University of Liverpool's Research Ethics Committee (approval number 11092). A mixed methods approach was utilised, including quantitative data gathered from the CJ assessment tool, focus groups with staff and students, and anonymised questionnaires for staff and students [10], [11].

The first stage of this work incorporated a comprehensive literature review using key databases to discover existing literature and identify gaps within the evidence base. Searches were undertaken utilising the following databases: Web of Science, Scopus, and Education Resources Information Centre.

Following this, the team's digital education expert developed the Comparative Judgment Application and integrated it into CANVAS (the university's virtual learning environment) so key personnel could access the tool. It was also decided that focus groups would be undertaken with the staff to evaluate the CJ application implementation process. The focus groups facilitated by an individual not involved in the CJ pilot, adding rigour to the research process. Emergent themes will be captured, and recommendations and future developments to the tool will be made.

Following the integration of the CJ application, the tool was piloted as a learning, assessment, and feedback tool for an undergraduate module with a poster assessment element. The CJ application ensured that all data was anonymised before analysis and subsequent publication of assessment results. Both students and staff used the pilot version of the 'CJ application', which generates a vast collection of feedback and a 'rank' of appraised poster submissions. These posters were also marked independently using the 'traditional' method, and analysis and evaluation were undertaken between three groups of data (the 'traditionally marked' group, the CJ student marked group, and the CJ staff marked group).

Focus groups were then undertaken with academic staff and students based on their overall experience of the CJ process. All participants' contributions were anonymised when generating transcripts, and all participants were reminded of the confidential nature of the research process.

Participant anonymity was maintained throughout the study via data collection, with individual identifiers removed and overall data grouped (for example, marks). All anonymised data (identifiers removed) was secured on a university server in a password-protected folder, with only the Principal Investigator and research team able to access this data.

3 RESULTS

The results from this preliminary study are currently being collated and analysed by the CJ research project team. We will present our findings and subsequent conclusions at the INTED conference in March 2024.

However, from the process thus far, certain conclusions can be drawn from observations, anecdotal evidence, and subsequent reflections. Following our literature review and current provision, we identified an area of CJ that could be improved. We saw the opportunity to develop a CJ tool primarily focused on learning with an emphasis on providing feedback without the distraction or the pressure of generating a grade. All too often, the attention is drawn to the 'better' of the two submissions when employing a CJ approach, with the 'lesser/ weaker' submission receiving less or no feedback. This approach ensures equity in the opportunities for assessor feedback provision. Assessors move through the same steps for both submissions in the comparison, giving each an equal opportunity for observation and critique.

4 CHALLENGES

We are providing enough structure to the feedback process to maintain a focus on the learning objectives without constraining the assessors' field of view'. An initial challenge was staff requesting detailed marking guidance and rubrics to structure the marking. There was some initial inertia to overcome in using an approach which required implementing an 'internal rubric/ guidance'.

We are reducing the task's complexity, focusing on the overarching objective, and keeping the process simple. In doing this, assessors and students can build confidence quickly, owing to the ease of the approach and its simplicity. They are articulating their thoughts concisely and constructively, simultaneously developing their knowledge and understanding of the topic.

When assessing work, it is all too easy to focus on the areas that are incorrect, or at the extreme end of a scale. However, this can result in the learner missing out on valuable insight into the good points of the work, or those areas which would benefit from a minor amendment or upgrade. We implemented a simple three-point, colour-coded focus to the feedback. This encouraged a balanced approach to providing feedback, encouraging the assessor to increase their engagement with each submission.

From observation, we have observed positive reflections from staff who have used the tool. Reporting anecdotally, staff have stated that there is an ease and simplicity by which they could quickly and efficiently provide feedback on many submissions. Students again anecdotally reported an increase in the amount and quality of feedback provided. In many cases, this was seen as a positive. However, in some submissions, the feedback initially appeared to demonstrate a conflict of opinion and highlighted the variability in academic judgement. Nonetheless, further, more detailed observation of the feedback and associated judgement scales clarified the consensus of the assessor's opinions in these instances.

5 QUANTITATIVE V QUALITATIVE JUDGEMENTS

Discussions with both the staff and the students recognised the value of feedback in the form of qualitative data. They provided academic insight and granularity to help develop a deeper understanding

of the submission's positives and areas for improvement. In addition, it became apparent that quantitative scoring in critical areas was also of great value. Initially, this took the form of three key aspects: knowledge, accessibility, and use of space (key objectives for a poster). Staff and students found this useful to put the written feedback into perspective. However, feedback on the process identified a potential area for development, increasing the number of areas in which this quantitative 'Likert' judgement could be expressed. The quantitative element allows the employment of 'parameter ranking' and potential grouping of 'similarly ranked' submissions for the more 'closely matched' submissions to be compared.

The simplicity of the low-stakes, high-yield CJ tool is such that it is easily transferable across disciplines. The language used in its qualitative and quantitative feedback guidance is universally applicable, thus opening limitless possibilities for implementation in learning and assessment.

6 CONCLUSION

In conclusion, the development and implementation of the Comparative Judgment (CJ) assessment tool in the context of the School of Life Sciences (SoLS) at the University of Liverpool have been driven by a commitment to enhancing the learning experience and feedback provision for students. The paper has outlined the rationale behind choosing CJ as a pedagogical strategy, emphasising its potential to foster fair, objective, and consistent assessment practices. The ongoing pilot project, aimed at developing a bespoke digital CJ tool, represents a significant step towards realising this vision.

The methodology section underscores the comprehensive and collaborative approach taken in this endeavour, involving students, staff, and digital experts in the tool's development, implementation, and refinement. Ethical considerations have been carefully addressed, ensuring the research is conducted with integrity and sensitivity to participant confidentiality.

Despite the paper being presented before the completion of the data analysis, the preliminary observations and reflections from the ongoing study have provided valuable insights. The decision to create a custom CJ tool focused on learning and feedback instead of traditional grading addresses a specific gap identified in the existing literature. The emphasis on equity in feedback provision for both compared submissions is noteworthy, aiming to ensure a balanced and constructive assessment process.

Challenges encountered during the project, such as providing sufficient structure to feedback without constraining assessors, overcoming initial resistance to internal rubrics, and maintaining simplicity in the process, have been transparently acknowledged. Strategies to address these challenges, including a three-point, colour-coded focus for feedback, have been outlined, showcasing the adaptability and responsiveness of the CJ tool.

The paper highlights positive feedback from staff and students who participated in the pilot project, with anecdotal evidence suggesting increased efficiency in providing feedback and a positive impact on students' engagement. The acknowledgement of the need for both qualitative and quantitative feedback, along with the flexibility of the CJ tool across disciplines, further underscores its potential as an innovative and universally applicable assessment strategy.

In summary, while the conclusions based on the complete data analysis are pending, the ongoing CJ research project at SoLS, University of Liverpool, appears promising. The commitment to refining and enhancing the assessment tool based on the feedback received reflects a continuous improvement mindset. The anticipation of presenting findings at the INTED 2024 conference underscores the commitment to sharing insights and contributing to the broader educational community.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the academic staff, digital education experts, and students who directly and indirectly contributed to this work.

REFERENCES

[1] E. J. Ashbaugh. "Reducing the variability of teachers' marks". *Journal of Educational Research*, vol. 9, pp. 185–198, 1924. Retrieved from: https://doi.org/10.1080/00220671.1924.10879447

[2] S.R. Bartholomew, G.J. Strimel, and E. Yoshikawa. "Using adaptive comparative judgment for student formative feedback and learning during a middle school design project", *International Journal of Technology and Design Education*, vol. 29, no. 2, pp. 363–385, 2019. Retrieved from: https://doi.org/10.1007/s10798-018-9442-7.

[3] T. Potter, T. *et al.* "ComPAIR: A New Online Tool Using Adaptive Comparative Judgement to Support Learning with Peer Feedback", *Teaching and Learning Inquiry*, vol. 5, no.2, pp. 89–113, 2017. Retrieved from: https://doi.org/10.20343/teachlearninqu.5.2.8.

[4] A. Pollitt. "The method of adaptive comparative judgment". *Assessment in Education: Principles, Policy & Practice,* vol. 19, no. 3, pp. 281–300, 2012b. Retrieved from: https://doi.org/10.1080/0969594X.2012.665354

[5] M. Lesterhuis. "The validity of comparative judgment for assessing text quality: An assessors' perspective"*. PhD Thesis*. University of Antwerp. 2018

[6] E. Hartell., and J. Buckley. "Comparative Judgment: An Overview". In: A. Marcus-Quinn, and T. Hourigan. (eds) *Handbook for Online Learning Contexts: Digital, Mobile and Open*. Springer, Cham. 2021. Retrieved from: https://doi.org/10.1007/978-3-030-67349-9_20

[7] A. Pollitt. "Comparative judgment for assessment". *International Journal of Technology and Design Education,* vol 22. no.2, pp. 157–170, 2012a. Retrieved from: https://doi.org/10.1007/s10798-011-9189-x

[8] W.W. Willingham, J.M. Pollack, and C. Lewis. "Grades and test scores: Accounting for observed differences". *Journal of Educational Measurement*, vol. 39, pp. 1–37, 2002. Retrieved from: doi.org/10.1002/j.2333-8504.2000.tb01838.x

[9] P.J. Mensink. and K. King. "Student access of online feedback is modified by the availability of assessment marks, gender and academic performance", *British Journal of Educational Technology*, vol.51, no. 1, pp. 10–22, 2020. Retrieved from: doi:10.1111/bjet.12752

[10] A. Divan., L.O. Ludwig, K.E. Matthews, P.M. Motley, and A.M. Tomljenovic-Berube. "Survey of Research Approaches Utilised in the Scholarship of Teaching and Learning Publications". *Teaching and Learning Inquiry,* vol. 5, no.2, pp. 16-29. 2017. Retrieved from: https://doi.org/10.20343/teachlearninqu.5.2.3.

[11] L.D, Goodwin and W.L. Goodwin. "Qualitative Vs. Quantitative Research or Qualitative and Quantitative Research?", *Nursing Research.* vol. 33, no. 6, pp. 378–384. 1984 Retrieved from: https://doi.org/10.1097/00006199-198411000-00022