Received: 17 March 2025 | Accepted: 3 March 2026

DOI: 10.1002/rev3.70149

RESEARCH REPORT

Review of Education BERA

Rendering transparency to ranking in educational assessment via Bayesian comparative judgement

Andy Gray[1,2] | **Stephen Lindsay**[3] | **Jen Pearson**[2] | **Tom Crick**[4] | **Alma Rahat**[5]

[1]School of Design, Bath Spa University, Bath, UK

[2]School of Mathmatics and Computer Science, Swansea University, Swansea, UK

[3]School of Computer Science, University of Glasgow, Glasgow, UK

[4]School of Education, University of Bristol, Bristol, UK

[5]Computer Science, Loughborough University, Loughborough, UK

Correspondence
Andy Gray, Swansea University, Fabian Way, Crymlyn Burrows, Skewen, Swansea SA1 8EN, UK; and Bath Spa University, Newton St Loe, Bath, BA2 9BN, UK.
Email: a.gray2@bathspa.ac.uk; 445348@swansea.ac.uk

Funding information
UK Research and Innovation

Abstract

Transparency in educational assessment has become an increasingly pressing concern, particularly in the aftermath of the pandemic, as institutions seek more equitable, robust and defensible methods of evaluating student work. Comparative judgement (CJ) has gained traction as a promising alternative to traditional rubric-based marking. However, despite its potential, CJ has been criticised for its perceived opacity, particularly in high-stakes contexts where fairness, auditability and trust are paramount. This paper investigates whether Bayesian comparative judgement (BCJ), which applies Bayesian statistical methods to CJ, can enhance transparency by making the judgement process more structured, interpretable and accountable. BCJ introduces a probabilistic framework that incorporates prior knowledge and updates beliefs based on new evidence, allowing for quantification of uncertainty and clearer justification of ranking decisions. It enables greater insight into the consistency of judgements and highlights areas of disagreement among assessors. We also evaluate a recent multi-criteria extension of BCJ that models each learning outcome (LO) separately, mirroring the structure of rubric-based assessment while retaining the efficiency and comparative strengths of CJ. This approach supports the generation of both outcome-specific and holistic rankings, offering detailed feedback without sacrificing the coherence of the overall

evaluation. Using real-world assessment data from a UK higher education course involving experienced professional markers, we demonstrate the application of BCJ and multi-criteria BCJ in practice. Our analysis highlights how these models can provide rigorous, transparent insights into the reasoning behind both individual and collective rankings. We also discuss how BCJ supports external validation of assessment outcomes. Finally, through semi-structured discussions with participant markers and expert CJ practitioners, we qualitatively assess the perceived transparency and usefulness of BCJ in authentic settings, particularly where high-stakes decisions are made. We conclude by outlining the benefits and limitations of BCJ and its relevance across varied educational contexts.

KEYWORDS

active learning, assessment, Bayesian statistics, comparative judgement, higher education, machine learning, transparency

Context and implications

Rationale for this study: Bayesian Comparative Judgement (BCJ) and its multi-criteria extension (MBCJ) enhance assessment transparency by explicitly quantifying uncertainty and assessor disagreement within ranking outcomes.

Why the new findings matter: Quantitative analysis using real higher-education assessment data demonstrates that BCJ and MBCJ produce interpretable and defensible rankings when compared with traditional marking practices.

Implications: Perspectives from professional markers and assessment experts indicate that BCJ supports fairness, trust, auditability and manageable workload in authentic high-stakes assessment contexts.

INTRODUCTION

The transparency of assessment practices in education is a significant concern, particularly in light of recent global shifts that underscore the need for fairer, more rigorous, and accountable assessment systems (Knight et al., 2025; Nisbet & Shaw, 2020)—perhaps more so following the COVID-19 pandemic (Crick, 2021; Watermeyer, Crick, et al., 2021), and the inexorable rise and widespread impact of AI in education (Dwivedi et al., 2021, 2023; Swiecki et al., 2022). In the UK, transparency challenges are exacerbated by the high workload pressures faced by teachers and academics, with assessment being one of the main contributors (Department for Education, 2024; Morris et al., 2023). Time pressure on marking can lead to inconsistencies, making fairness and transparency in educational assessment paramount (Rasooli et al., 2018; Tierney, 2014).

Teachers in England work an average of 50 h per week, while school leaders work more than 55, according to the UK Government's Department for Education (DfE)'s workload survey (Department for Education, 2019) (*N.B*. the DfE has responsibility for education in England only, due to devolved education policy in the UK across the four nations). Time spent on unnecessary tasks driven by an accountability regime contributes to the ongoing recruitment and retention crisis without improving learning outcomes (LOs) (National Education Union, 2024). Recognising this, the DfE has emphasised the need to reduce school workload, providing a toolkit of practical resources to support reductions (Department for Education, 2022).

In 2019, the average self-reported working hours for all teachers and middle leaders was 49.5 h per week, a reduction of 4.9 h from 2016. Primary teachers and middle leaders reported working an average of 50 h per week in 2019, down from 55.5 in 2016, while secondary teachers and middle leaders reported a decrease from 53.5 h to 49.1 (Department for Education, 2019). However, primary teachers continue to work longer hours than their secondary counterparts, though the gap has narrowed from 2 h per week in 2016 to 0.9 h in 2019. Clearly, there is a drive to improve working conditions across the UK, and arguably it has had some positive impact.

Nonetheless, marking remains one of the most time-consuming tasks. According to the DfE's workload survey, 61% of secondary school teachers and middle leaders reported that they spend too much time marking (Department for Education, 2019, 2024). The proportion of teachers who feel overwhelmed by marking has remained persistently high, with 43% reporting excessive marking workloads in 2024, compared to 46% in both 2022 and 2023 (Department for Education, 2024).

This marking workload leads to stress, poor wellbeing (Jerrim & Sims, 2021), and has also been shown to be associated with systematic shifts in grading behaviour under repetitive marking conditions (Erturk et al., 2022). As student numbers increase, absolute marking becomes increasingly difficult to sustain at scale, motivating interest in automated approaches (Senanayake & Asanka, 2024). Furthermore, there is a disconnect between grading policy and practice, as many teachers do not consistently use written criteria, often relying on holistic rather than analytical judgements (Bloxham et al. 2011), which can contribute to variability in grading outcomes across educational contexts (Hausdorff & Farr, 1965).

While these workload challenges are well-documented in schools, they are mirrored in UK higher education (HE) institutions (and indeed, internationally), where marking burdens are similarly acute. Academics face growing class sizes, expanded assessment formats, and intensifying institutional accountability demands, with the resulting impact on their health and wellbeing (Hardman et al., 2022; McGaughey et al., 2022; Watermeyer, Shankar, et al., 2021). As a result, HE staff operate under comparable marking pressures, with heavy workloads, compressed timelines, and overlapping modular deadlines creating clear conditions for marking fatigue. These structural constraints are widely recognised as placing strain on the consistency and quality of assessment (Norton et al., 2019; Raaper, 2016; Spencer & Horn, 2023).

These HE sector assessment transparency issues are compounded by structural sector-wide issues in part arising from the COVID-19 pandemic (Watermeyer et al., 2022, 2025; Watermeyer, Shankar, et al., 2021). Academic staff face increasing workload and thus marking burdens due to growing class sizes, diverse assessment formats, and institutional pressure to provide detailed feedback within tight timeframes (Raaper, 2016; Siegel et al., 2021; Spencer & Horn, 2023). The marking of more traditional assessments, such as essays, project reports, and reflective pieces, requires considerable time investment, particularly when attempting to apply complex rubrics consistently across a cohort. These challenges are intensified in modular structures where assessments are frequent and staggered, leading to multiple overlapping marking deadlines and feedback cycles throughout the academic year (Norton et al., 2019). Under these conditions of cognitive fatigue and time pressure, the reliability and fairness of assessment can potentially be compromised (Hasan & Jones, 2024).

In UK HE, assessment transparency has become a core component of institutional accountability, quality assurance, and external scrutiny (Norton & Hack, 2024). The Office for Students (OfS), Quality Assurance Agency (QAA), and Advance HE all emphasise the need for defensible, equitable assessment practices (Norton & Hack, 2024; Office for Students, 2022; Quality Assurance Agency for Higher Education, 2018). These policy imperatives have intensified post-pandemic, as institutions recalibrate assessment models for resilience, fairness, and transparency (Walker, 2025). Thus, despite sector-wide initiatives to enhance and improve efficiency of assessment in UK HE, such as automated grading for objective assessments and standardised rubrics, many assessments still rely on human judgement for evaluating higher-order skills. Consequently, inconsistencies in grading remain a concern, particularly when assessments are distributed between multiple markers (Ragolane et al., 2024). As the frequency and volume of marking increase, opportunities for moderation and second marking become limited, further reducing transparency and perceived fairness. These issues have prompted calls for new approaches that can reduce marker burden while preserving or enhancing the validity and rigour of assessment processes (Bloxham, 2009). However, new and innovative methods must be understood in relation to these governance frameworks that shape their adoption and perceived legitimacy.

While rubric-based grading is widely used to promote transparency and consistency, its effectiveness in addressing subjective inconsistencies and capturing nuanced student performance depends heavily on design and implementation (Velasco-Martinez & Tojar-Hurtado, 2019). Comparative judgement (CJ) has emerged as an alternative method with the potential to overcome some of these limitations, offering a ranking system based on direct—pairwise—comparisons of student work rather than pre-defined scoring criteria (Jones & Davies, 2024). However, CJ has been described by some stakeholders as opaque, since assessor decisions are aggregated statistically and individual rationales are not readily visible, raising questions about transparency (Holmes et al., 2020). To address these concerns, we explore the application of Bayesian CJ (BCJ) to render the CJ process more transparent and interpretable. This study thus situates itself squarely within the UK HE context, using real-world assessment data from a postgraduate course to evaluate the applicability and impact of BCJ and MBCJ in HE-specific assessment scenarios. While the broader motivation draws from systemic concerns about assessment fairness and workload, the findings offer targeted implications for academic staff and institutional policy in HE.

BCJ leverages a Bayesian probabilistic framework. Gray et al. (2024) demonstrate how prior knowledge about each pairwise judgement can be elicited and how the posterior can be updated as new evidence becomes available—specifically, when an assessor decides the winner in a comparison between two pieces of student work (or items). The uncertainty around pairwise comparisons can then be propagated to generate a distribution over the rank order of individual items, which can subsequently be converted into letter grades using the cumulative density associated with each item's rank distribution.

In their follow-up work, Gray et al. (2025) introduce methods for evaluating the reliability and consistency of the ranking process using novel metrics. These include *Rank Separation Reliability* (RSR), which assesses how effectively the ranking procedure differentiates item ranks at a global level, and the *Mode Agreement Percentage* (MAP) and *Expected Agreement Percentage* (EAP), which quantify agreement on a per-pair basis and help identify pairs that divide opinion among assessors—capturing both intra- and inter-assessor consistency. Estimating MAP or EAP further enables straightforward identification of contentious pairs, allowing a chief assessor to intervene and adjudicate where necessary.

Additionally, their work demonstrates how multiple criteria or learning outcomes (LOs) can be incorporated into pairwise comparisons, yielding both LO-specific and overall rank distributions for each item. This represents the first principled approach to multi-criteria BCJ (MBCJ) where LO-specific weights are known. The framework enables detailed,

criterion-specific assessment—covering dimensions such as critical thinking, technical proficiency, and creativity—thereby producing granular insights similar to rubric-based assessment, while preserving the efficiency of pairwise comparison. MBCJ also generates a holistic overall ranking that synthesises information across LO models, aligning with multi-criteria rubric-based approaches (Gray et al., 2025).

Both BCJ and MBCJ are adaptive frameworks that employ active-learning strategies based on entropy, selecting at each step the pair with the greatest uncertainty— and therefore the most informative—for assessor judgement.

In contrast, traditional CJ derives item scores on a latent scale that may not correspond to any interpretable or meaningful notion of a 'true' score. These scores are then used to rank items, but under the standard frequentist maximum-likelihood Bradley–Terry Model (BTM), only point-estimates are available. In Bayesian BTM (BBTM), a prior is imposed over item scores and uncertainty is estimated using Markov Chain Monte Carlo (MCMC) methods, which are computationally intensive and can be impractical for large item sets Wainer (2023). Moreover, both BTM and BBTM suffer from limited transparency: the inferred scores may not reflect meaningful values, and areas of assessor disagreement are obscured.

BCJ offers several advantages over standard CJ. First, BCJ exhibits superior convergence properties compared with BTM (Gray et al., 2024): it achieves accurate rankings using fewer comparisons and continues to improve as additional evidence accumulates, whereas BTM performance may deteriorate beyond a certain threshold. Second, BCJ's adaptive learning strategy provides consistent improvements over common pair-selection methods such as random or round-robin selection. Gray et al. did not compare against Pollitt's adaptive CJ (Pollitt, 2012) due to concerns about its reliability (Bramley, 2015). Third, BCJ enhances transparency by enabling direct auditing of pairwise comparisons, visualising uncertainty in decision-making, and quantifying agreement through EAP. In contrast, traditional CJ typically provides only a single maximum-likelihood estimate of the probability of a winner per pair, which may be overconfident and fails to capture the full structure of assessor disagreement under uncertainty—see, for instance, the mode agreement percentage (MAP) metric proposed by Gray et al. (2025), which is essentially equivalent to this. BCJ's rank distributions are directly interpretable as order statistics rather than latent-scale scores, and EAP enables a chief assessor to retrospectively review and revise problematic decisions—an innovation absent from traditional CJ. Crucially, BTM cannot accommodate multi-criteria extensions, whereas MBCJ can.

For the first time, this paper demonstrates the effectiveness of BCJ in real-world educational settings by employing a genuine validation dataset comprising officially marked assignments (via *absolute value judgements*) from a postgraduate course at a UK higher-education institution. This dataset allows us to illustrate the practical implementation of BCJ and MBCJ, and to evaluate their impact on transparency, fairness, and accountability in assessment. By integrating quantitative analyses with qualitative insights from professional markers and expert commentators, we assess the extent to which BCJ and MBCJ provide clear and defensible rationales for both individual and aggregate judgements. Furthermore, we examine their potential to identify ambiguity or conflict in assessment decisions, particularly in high-stakes contexts—such as national assessments—that require heightened transparency (Holmes et al., 2020). Therefore, the main contributions of this study are as follows:

- Illustration of the improvement in transparency BCJ offers over CJ by offering a structured process that tracks and explains decision-making, and providing estimations of uncertainty in rankings.
- Evaluation and comparison of traditional, BCJ and MBCJ to standard marking within a real-world assessment context.

- Analyses of insights from educators' experiences from a UK HE context—through a combination of quantitative data analyses and discussions with professional markers and experts in CJ showing how educators perceive these approaches in terms of fairness, workload, and usefulness.

We review the literature and background in "Literature and background" section; "Experimental settings" section looks at the experiment's setup and the methodologies used; "Results and discussion" section looks at the results and discusses them; while in "Conclusions" section, we make our conclusions about the study.

LITERATURE AND BACKGROUND

In this section, we explore the importance of assessments within education and the concerns about traditional marking. Furthermore, we review the literature on the rapidly growing approach to assessments called CJ. Finally, we explain the process of BCJ and we explore the multi-criteria approach to BCJ or MBCJ.

Traditional educational assessment methods

Traditional marking where we assign a value or score to a piece of work under assessment is the dominant form of grading in education. Teachers assign marks based on fixed criteria or rubrics, aiming to gauge a student's performance in an absolute sense. Despite widespread use, this approach has been critiqued for issues related to consistency, bias, transparency, and, crucially, the cognitive demands it places on educators.

Biases—based on a student's previous performance, personality or even handwriting—can affect marks. Scharaschkin and Baird (2000) highlights the 'halo effect', where a teacher's perception of a student's past achievements influences their grading of current work. This can lead to an over- or underestimation of a student's true capabilities, which is particularly problematic in high-stakes assessments. Even with a grading rubric, different teachers can interpret the same criteria in varying ways, leading to discrepancies in scoring (de Moira et al., 2002; Scharaschkin & Baird, 2000). Biases also stem from factors like teacher fatigue, stress, and subjective preferences. Grading decisions are often influenced by non-academic factors, even subconsciously. For example, assessors may give higher marks to work that aligns more closely with their own views or personal standards (Willey & Gardner, 2010). These undermine the fairness of assessments, affecting students' opportunities and their trust in the educational system (Guskey, 2024; Read et al., 2005).

While anonymisation is often deemed as a definitive way to combat biases and improve trust in the system, it may not necessarily be effective (see, for example, Pitt and Winstone (2018)) and is not always feasible, for instance, for presentations. In CJ, since assessors only view pairs in isolation and are less likely to let factors unrelated to quality, such as familiarity with a topic or presentation style, sway their judgment (Mentzer et al., 2021).

Transparency in assessment refers to the clarity and openness with which grading criteria and decisions are communicated to students and other stakeholders. Traditional marking often lacks this transparency, as students may receive little feedback beyond an overall numerical score or grade, leaving them unclear about the aspects of their work needing improvement. This hinders learning processes, as students cannot fully understand how they are evaluated or how they can improve (Bamber, 2015).

Transparency is vital to ensuring trust in the assessment process, yet traditional marking methods often fall short in providing the necessary clarity (Ilahi et al., 2024). When

grading is inconsistent or biased, students and parents may feel frustrated or sceptical about the fairness of the assessment process. This lack of transparency is especially concerning in high-stakes situations, where grades significantly impact future educational and career opportunities.

Traditional marking is cognitively demanding for teachers. Drawing on cognitive load theory, the mental resources required to assess large volumes of work accurately and consistently can become depleted over time. Evidence suggests that such cognitive strain can lead to marking fatigue, with grading quality deteriorating as assessors progress through a sequence of assessments (Hasan & Jones, 2024). These risks are further exacerbated by performance pressures and accountability demands within contemporary education systems (Grissom et al., 2015). This depletion is even more pronounced when teachers are required to mark open-ended or complex tasks, as these assessments require continuous decision-making and interpretation (Pollitt, 2012).

In terms of actual time spent, teachers would often spend hours outside of class reviewing and grading work in traditional marking (Morris et al., 2023). This not only impacts their workload, but may also limit the time they have available for other important teaching activities, such as lesson planning and providing one-on-one support to students (Jerrim & Sims, 2021). Teachers operate under significant performance and accountability pressures that constrain professional practice (Grissom et al., 2015). In assessment contexts, judgement processes are known to rely on holistic impressions and implicit standards rather than fully analytic evaluation (Bramley, 2015). Under such conditions, time pressure may increase the likelihood that assessors draw on these judgement heuristics when grading. These shortcuts, while understandable, compromise the reliability and validity of assessments.

Traditional absolute marking in education presents several challenges, including inconsistencies, bias (Magowan, 2023), a lack of transparency (Gonsalves & Lin, 2025), and the negative impact on teachers' wellbeing (Jerrim & Sims, 2021). These issues highlight the need for more reliable, transparent, and efficient assessment methods that support both educators and students in the learning process.

Comparative judgement

CJ has emerged as a promising approach in educational assessment, offering an alternative to traditional scoring methods for evaluating complex, subjective tasks. Rather than scoring individual pieces of work with a numerical grade or rubric, CJ relies on the collective judgements of experts who perform pairwise comparisons of student work and decide which better represents the LO. CJ has demonstrated advantages such as reducing subjective bias and achieving higher reliability in assessments compared to traditional methods (Crompvoets et al., 2022; Pollitt, 2012).

The conceptual roots of CJ date back to Thurstone's Law of CJ (Thurstone, 1927), which posits that humans are more consistent in making comparative rather than absolute judgements. In educational assessment, this model has been increasingly adopted because it capitalises on humans' relative strength in making comparative rather than absolute judgements, enabling more accurate evaluations without reliance on pre-defined criteria or rubrics (Bramley, 2015).

In fact, Laming et al. argued that all human judgements are fundamentally relative in nature (Laming, 1984), a claim that is frequently cited as a key rationale for the use of CJ. Moreover, people tend to make relative judgements more quickly (Mussweiler & Epstude, 2009), and there is evidence that such judgements can reduce cognitive load under certain conditions (Palisse et al., 2023). Assessors may also find it easier to decide which of two pieces of work is better than to evaluate a single piece of work against a

complex rubric (Goossens & De Maeyer, 2017). However, relative judgements are not universally superior: they do not always outperform absolute judgements, and their advantages may depend on context (Kelly et al., 2022).

CJ as a structured method in education was popularised by Pollitt (2012), who introduced a software platform allowing assessors to make pairwise comparisons systematically. This enables the aggregation of judgements to produce a ranking order, where each student's work is placed relative to others based on the majority of comparisons made by judges. This method offers an innovative, user-friendly interface for assessors and facilitates faster and more reliable marking of complex work.

The overall CJ process is summarised in Figure 1. We would start with all the items that are to be ranked, and then select a pair of items that must be shown to the assessors for determining the winner. We then deploy an appropriate statistical method to determine the overall rank, which would be reported at the end of the process, once we go beyond a specified threshold for the number of pairs to be shown to the assessors.

The different variants of CJ considered in this paper vary in how pairs are selected, how winners are determined, and how overall ranks are produced. In standard (non-adaptive) CJ, the most common approach to pair selection is uniform random sampling from the set of all possible pairs (Jones & Davies, 2024). While it is possible for an automated process (e.g., a large language model (Gu et al., 2024)) to determine the winner in a paired comparison, in our context we typically observe that one or more humans make this decision synchronously or asynchronously. Once the winner is known and the current dataset is augmented with this new information, we use a statistical method to derive the overall rank. In traditional CJ, we use Bradley-Terry model (BTM) to generate the overall rank. In BTM, it is assumed that there is a likelihood over the score of an individual item, and then some approach towards maximising the likelihoods across all items, for example, the minorisation–maximisation method (Hunter, 2004), locates the optimal parameter set. The corresponding score to the optimal

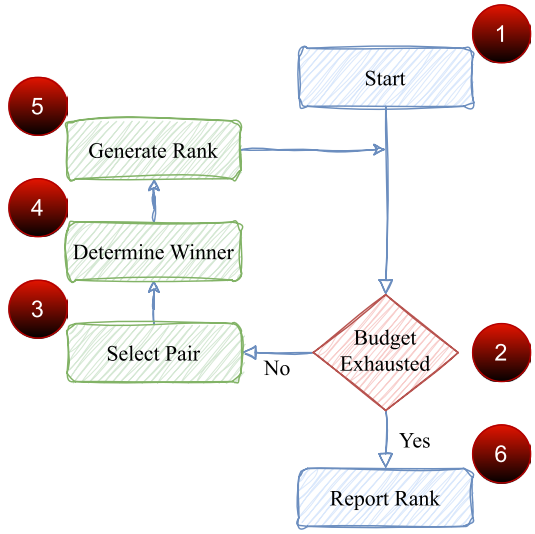

FIGURE 1 Flowchart depicting the CJ process. We start with a number of items to rank in Step 1. Then, based on the budget on how many pairs we can show the assessors (Step 2), we firstly select a pair to show using an appropriate method (Step 3). Then the assessor would pick the winner (Step 4), and the statistical method in place would generate a rank for all the items in light of new evidence (Step 5). Once the budget is exhausted, we would report the final rank to the assessment owner (Step 6). The green boxes are the core elements of CJ that vary methodologically between distinct approaches (e.g., BTM, BCJ or MBCJ).

parameter set can be used to rank the items, all the way from 1 to N, where the highest score is ranked 1, and so on.

CJ has several advantages over traditional assessment methods. First, it has been shown to increase inter-rater reliability, meaning different assessors are more likely to agree on the ranking of work, even without detailed rubrics. For instance, literature has found that CJ produced more reliable assessments for open-ended tasks, such as essays and projects (Bramley & Vitello, 2019; Jones & Inglis, 2015), where defining a standard scoring rubric is challenging (Steedle & Ferrara, 2016). Another potential benefit of CJ is improved efficiency in assessment design and execution. Rather than developing and training assessors to apply detailed analytic rubrics, CJ relies on relatively quick pairwise comparisons, which can reduce assessor preparation and scoring complexity (Steedle & Ferrara, 2016). Assessors often find it easier to determine which of two pieces of work is better than to evaluate individual work against a complex rubric (Goossens & De Maeyer, 2017). Steedle and Ferrara (2016) highlight this, showing that teachers assessing writing tasks were able to reach reliable consensus more efficiently using CJ than through traditional rubric-based approaches. Moreover, CJ has been shown to reduce the influence of individual rater biases that can affect traditional scoring, as judgements are based on repeated pairwise comparisons and aggregated across assessors, limiting the impact of factors unrelated to the quality of the work (Mentzer et al., 2021).

While CJ offers clear advantages, it has been noted that its outcomes can be less immediately transparent to stakeholders than those produced through traditional grading approaches (Holmes et al., 2020). When used in isolation, CJ produces holistic judgements rather than explicit criterion-level feedback, which can make it harder for students and teachers to identify specific areas for improvement (Stuulen et al., 2024). This gap can be addressed by MBCJ (Gray et al., 2025).

In CJ, the number of decisions that assessors can feasibly make is often small relative to the total number of possible item pairs, since the latter grows combinatorially with the number of items N. The number of unique unordered pairs is

$$z = \frac{N(N-1)}{2} \tag{1}$$

For example, when $N = 10$, there are $z = 45$ distinct pairs. If we follow the common heuristic of collecting $10N = 100$ comparisons, this yields on average just over two decisions per pair, with some pairs typically receiving more than others.

However, when $N = 100$, the number of unique pairs increases dramatically to $z = 4950$. Under the same $10N$ heuristic (i.e., 1000 comparisons), the average number of decisions per pair falls below one. Thus, as the number of items increases, the number of potential pairs expands rapidly while the feasible number of comparisons remains modest.

For practical reasons—and informed by empirical reliability studies—it is often recommended to conduct approximately $10N$ comparisons to avoid overburdening assessors while still achieving high reliability, as supported by a meta-analysis of 101 CJ datasets (Kinnear et al., 2025). Nonetheless, due to the combinatorial explosion in the number of pairs, this fixed comparison budget provides progressively poorer coverage per pair as N grows.

In recent years, CJ has attracted attention as a tool for assessing complex competencies, such as critical thinking and creativity, which are difficult to measure with standard tests (Bramley & Vitello, 2019). Researchers are also exploring automated and semi-automated CJ systems, which could reduce the burden on human assessors and enable more widespread use (Christodoulou, 2025).

CJ offers a compelling alternative to traditional assessment methods, especially for evaluating complex, open-ended tasks. Although it poses some challenges in terms of feedback and scalability, the method has demonstrated clear benefits in terms of reliability, efficiency,

and fairness. As research and technology in this area advance, CJ is increasingly becoming a valuable tool in the landscape of educational assessment, potentially reshaping how teachers, students, and policymakers view assessment practices.

Mathematical formulation

In this section, we provide a brief description of the standard CJ process as modelled using the BTM.

Consider a set of N items, $I = \{I_1, \ldots, I_N\}$. In BTM, each item is associated with a *latent* (i.e., unobserved) positive score, represented by the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)^{\mathsf{T}}$. Under this model, the probability that item I_i is preferred over item I_j is given by

$$P(i \succ j) = \frac{\gamma_i}{\gamma_i + \gamma_j}.$$

Assuming independence of pairwise outcomes, the log-likelihood of the performance vector $\boldsymbol{\gamma}$ is

$$L(\boldsymbol{\gamma}) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \left[\omega_{[i,j]} \ln(\gamma_i) - \omega_{[i,j]} \ln(\gamma_i + \gamma_j) \right], \tag{2}$$

where $\omega_{[i,j]}$ denotes the number of times item i won a pairwise comparison against item j. An iterative *minorisation–maximisation* (MM) algorithm is then used to estimate the optimal parameter values for this model (Hunter, 2004).

Once the latent scores have been estimated, the expected rank for each item is obtained by sorting $\boldsymbol{\gamma}$:

$$\boldsymbol{r}_{i \in I} = (N + 1) - \text{argsort}(\boldsymbol{\gamma}),$$

where rank 1 corresponds to the highest-scoring item. This corresponds to step 5 in the flow chart presented in Figure 1.

It should be noted that ranks provide a more natural and intuitive interpretation of outcomes derived from pairwise comparisons, especially when contrasted with latent scores whose numerical values may bear little or no meaningful relationship to true score scales.

As discussed earlier, the pair-selection method in step 3 is typically either *random* or *round-robin* (a.k.a. no-repeating pairs).

It should be noted that, in BTM-based CJ, the primary method for uncertainty estimation is to compute the standard error (SE) of the estimated latent scores. These SEs are then used to derive the reliability (often referred to as consistency) metric *Scale Separation Reliability* (SSR) (Kinnear et al., 2025). A key limitation is that accurate SE estimation typically requires a substantial number of decisions—often many tens per item (Crompvoets et al., 2022). From a practical standpoint, a rule of thumb of approximately 10 decisions per item has been suggested (Kinnear et al., 2025).

It should also be noted that, in a Bayesian treatment of the BTM (BBTM), priors are placed on the latent scores $\boldsymbol{\gamma}$, yielding posterior distributions over the same parameters (Wainer, 2023). However, the resulting posterior has no closed-form solution, so computationally intensive MCMC methods are typically required to approximate these distributions.

Nonetheless, the Bayesian approach offers several advantages over standard frequentist maximum-likelihood estimation: most importantly, it yields a *posterior distribution* rather than a single point estimate. This enables more informative quantification of uncertainty around item scores and can mitigate issues such as the removal of items that always lose or always win against every opponent (Caron & Doucet, 2012). However, the interpretability challenges associated with the latent score scale remain.

Bayesian comparative judgement

BCJ is a novel approach to assessment that leverages Bayesian statistical machine learning methods to improve the reliability and transparency of CJ methods. Bayesian approaches are an area of growing interest within CJ research, allowing for adaptive testing methods that can optimise the number of comparisons required by prioritising pairs that will yield the most useful information about relative performance (Gray et al., 2024). This approach has shown promise in improving the efficiency and accuracy of CJ assessments, particularly in education settings where resources are limited.

Firstly, we introduce intuitively how the Bayesian statistics work; interested readers should consult Lambert (2018)'s book for an accessible treatment of the topic. BCJ starts by defining a model for a process under scrutiny. The model should have at least one parameter that controls the behaviour of the model, but can have many parameters that can be changed. We then impose, based on experience or knowledge of the system, some prior probability density over the parameter(s). As we collect data, we update the belief in the light of evidence, and produce an informed conclusion, that is, known as the posterior density over the parameter(s). The more data we collect, the more accurate posterior density becomes, and the uncertainty diminishes. The posterior is, at any point, our best guess given the data, and we use that to make a judgement about the process model parameters, and consequently about the process outcome.

This is best appreciated through an example of a biased coin. The model we can assume is one which produces either heads (e.g., success/win) or tails (e.g., failure/loss) given a certain probability, and can be controlled by a bias parameter that controls the probability of observing heads. For example, if the bias is 0.3, that means we would observe heads around 30% of the flips, and the rest would be tails. This type of biased-outcome parameter is often termed as a Bernoulli (random) variable, and the posterior is known to be a Beta density which can be readily updated via the Bayes' theorem. As we collect more data, the posterior density becomes more confident, that is, the uncertainty in the density reduces, and the most likely value starts to converge to the true bias. See Figure 2 for an example for 50 coin flips for a biased coin with weighting 0.3.

Unlike BBTM—where priors and posteriors are placed on latent scores γ—Gray et al. (2024) propose an alternative Bayesian approach to CJ: *BCJ*. In BCJ, each pairwise comparison is modelled directly as a Bernoulli outcome (i.e., a bias towards one item over another), making each comparison analogous to a coin flip and enabling Bayesian updating at the *pair* level. Furthermore, this update is computationally efficient because Bernoulli outcomes admit a conjugate Beta posterior, allowing closed-form Bayesian updating. This captures uncertainty for each comparison explicitly and provides clearer insight into per-pair agreement among judges. While BTM reports only a single point estimate of the pairwise probability, which can be misleading with few observations, BCJ yields a full Beta posterior. Its mode may coincide with the BTM estimate, but BCJ additionally quantifies the *width* (i.e., uncertainty) of the distribution—information absent from BTM's point output.

BCJ then derives *probabilistic ranks* for each item by aggregating the pairwise posteriors across all item pairs (under an independence assumption for comparison outcomes). An

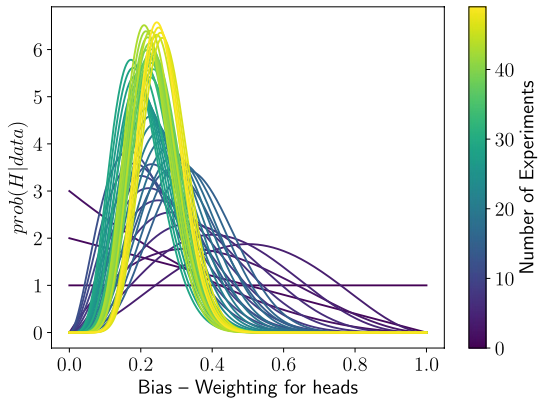

FIGURE 2 Illustration of Bayesian updates for a biased coin. The model is a generator for coin flipping outcomes: It will produce heads with the probability specified in the bias. Here, the bias is 0.3 (or 30% chance of observing heads) and we collected data for 50 (simulated) coin flips and updated the prior belief to track the posterior density. Without any observations, the horizontal line at 1 depicts the flat prior belief that the bias could be anything. As we collect more data, the posterior density—a Beta distribution for the Bernoulli bias variable—over the bias narrows, i.e., gets confident about the estimation, with a mode around 0.3; the lighter colours are later estimates of the density. The illustration was inspired from the work of Sivia and Skilling (2006).

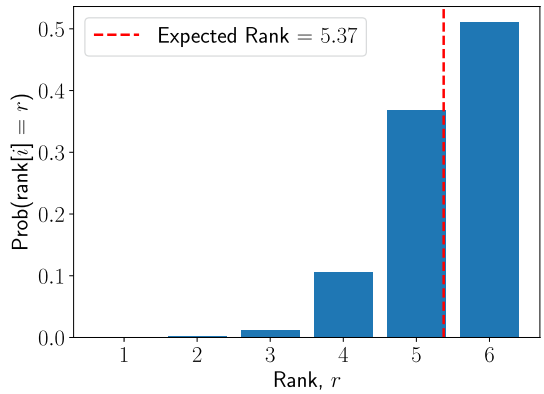

FIGURE 3 An example of rank density for an item *i* post BCJ, given 6 items. Here, this item has the highest probability (of around 50%) of being ranked 6, but the average (or expected) rank (shown in red dashed vertical line) is around 5.37 due to the consideration of uncertainty arising from paucity of data. BCJ uses the expected ranks (instead of scores in CJ) to determine the final ranks.

example rank-density distribution is shown in Figure 3, illustrating how an item is probabilistically distinguished and how its expected rank contributes to the overall ordering. For pair selection, Gray et al. (2024) introduce an entropy-based active-learning strategy: select the pair with the highest posterior entropy (conceptually proportional to the posterior width in Figure 2), that is, the pair whose outcome is most uncertain and therefore most informative. Across synthetic and real datasets, they demonstrate that BCJ paired with entropy-based active learning outperforms traditional BTM-based CJ in ranking accuracy, even when BTM is given substantially more judgements. A notable advantage is thus *interaction efficiency*: BCJ achieves higher accuracy with fewer comparisons, substantially reducing assessor workload.

BCJ's uncertainty estimates also highlight *contentious pairs*—cases where the posterior preference for *i* over *j* lies near 0.5 (e.g., within [0.25, 0.75]), indicating substantial disagreement among judges. Using the Expected Agreement Percentage (EAP)

introduced by Gray et al. (2024), such pairs can be flagged for escalation to a more experienced assessor. This targeted adjudication supports more nuanced decisions where traditional methods may struggle to separate items of similar quality, yielding results that are both more accurate and more trustworthy for complex assessments such as essays or project-based work.

In this context, BCJ's efficiency and adaptability make it feasible for applications in education, where resources and time are often constrained (Gray et al., 2024), while also providing output rank distributions that are more interpretable than latent scores.

Despite its advantages, BCJ systems can be complex to set up and require computational resources to calculate Bayesian probabilities in real time. While this has not been a barrier in experimental studies with a smaller number of items to compare, it could pose challenges for schools or smaller educational institutions with limited technology support. That is why a Monte Carlo (MC) version of the BCJ approach was also presented in Gray et al. (2024); we expand on potential implementation considerations in "Implementing BCJ" section.

Mathematical formulation

Following Gray et al. (2024), each pairwise decision is treated as a Bernoulli trial. For n comparisons of items i and j, with w wins for i and decisions $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ where $x_k \in \{0, 1\}$ indicates whether i beat j, the likelihood for the win probability p is

$$L(p|\boldsymbol{x}) = p^w (1-p)^{n-w}. \tag{3}$$

Using the conjugate Beta prior $p \sim \mathcal{B}(\alpha_{\text{prior}}, \beta_{\text{prior}})$ (uniform when $\alpha_{\text{prior}} = \beta_{\text{prior}} = 1$), the posterior after n comparisons is

$$\alpha_{\text{post}} = \alpha_{\text{prior}} + w, \qquad \beta_{\text{post}} = \beta_{\text{prior}} + (n - w),$$

so that the probability density function can be expressed as:

$$\pi(i \succ j) \equiv p \,|\, \boldsymbol{x} \sim \mathcal{B}(\alpha_{\text{post}}, \beta_{\text{post}}).$$

This probability density, along with its update under new evidence, is illustrated with a toy example in Figure 2.

Thus, the probability that item I_i is preferred over item I_j is

$$P(i \succ j) = P(p > 0.5) = 1 - \mathcal{F}_{\mathcal{B}}(0.5|\alpha_{\text{post}}, \beta_{\text{post}}) \tag{4}$$

where $\mathcal{F}_{\mathcal{B}}(\cdot)$ denotes the cumulative distribution function of the Beta distribution.

Gray et al. (2024) also derive a discrete distribution $P(r_i = a)$ for the rank of item i (details omitted, though an illustration is provided in Figure 3). The expected rank is then given by

$$\mathbb{E}[r_i] = \sum_{a=1}^{N} a P(r_i = a),$$

and can be used to obtain the final output ranks:

$$\boldsymbol{r}_{i \in I} = (N + 1) - \text{argsort}(\mathbb{E}[\boldsymbol{r}]).$$

For pair selection (Step 3 in Figure 1), they adopt uncertainty sampling: choose the pair with the highest entropy of its Beta posterior. The entropy of $\mathcal{B}(\alpha_{\text{post}}, \beta_{\text{post}})$ is (Lazo & Rathie, 1978)

$$H\left[\pi(i \succ j)\right] = \ln \mathcal{B}(\alpha_{post}, \beta_{post}) - (\alpha_{post} - 1)\psi(\alpha_{post}) - (\beta_{post} - 1)\psi(\beta_{post}) + (\alpha_{post} + \beta_{post} - 2)\psi(\alpha_{post} + \beta_{post}), \tag{5}$$

where $\mathcal{B}(\cdot, \cdot)$ is the Beta function and $\psi(\cdot)$ the digamma function. This strategy outperforms random and no-repeating-pairs baselines, and follows standard active learning principles (Lewis, 1995).

With this, the pair to be shown can be formally identified as:

$$(i, j) \leftarrow \arg \max_{\forall i,j \in \{1,\ldots,N\} \wedge i \neq j} H\left[\pi(i \succ j)\right].$$

Multi-criteria Bayesian comparative judgement

As discussed previously, CJ has been criticised for not considering multiple dimensions of comparisons: this reduces the richness of information for both students and assessors. Despite the accuracy conferred by the approach, this is an important shortcoming to note. BCJ also suffers from this; hence, traditional marking has a clear advantage in this aspect because judges consider the different areas of the rubric with an overall grade derived as a weighted sum of components, and thus, complete information is available and can be queried.

Gray et al. (2025) explored the idea of creating a multi-criteria version of BCJ for the first time. In their approach, pairwise comparison is performed in each individual component (or criterion) within a rubric, but at a time across all components for a pair that is being evaluated. This allows component-specific estimation of ranks per item and the probability densities therein. They then show how to combine them via weighted sum of the component cumulative densities and produce an overall rank density. Like BCJ, the expected ranks of items can be used to derive overall rankings. In addition, they also proposed an extension to the pair selection method based on combined entropy—again an estimation of the maximum utility of the next pair to be shown to the participant assessors. An illustration of the outcome expected ranks from the process is shown in Figure 4. Their experiments show that MBCJ is equivalent to BCJ in performance, without the loss of rich component-wise comparisons. In addition, they proposed a novel reliability measure for both BCJ and MBCJ.

Mathematical formulation

Considering D LOs and given the CDF of the preference distribution $\mathcal{F}_d(i \succ j)$ for the dth LO, the overall preference CDF for item i over item j can be expressed as a weighted mixture Lindsay (Lindsay, 1995):

$$\mathcal{F}(i \succ j) = \sum_{d=1}^{D} \lambda_d \mathcal{F}_d(i \succ j), \tag{6}$$

where λ_d denotes the contribution of LO_d to the overall mark. This mixture can then be used directly to compute the overall probability of preference between items (analogous to Equation (4)), from which the rank distribution follows (Step 5 in Figure 1).

For pair selection (Step 3 in Figure 1), assuming independence across LOs, the total entropy across all D LOs is computed as Korn and Korn (2000):

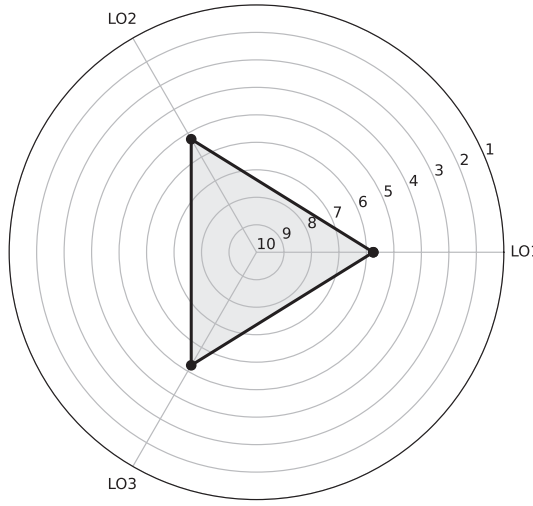

FIGURE 4 Radar plot depicting an item's expected rank $\mathbb{E}[r]$ (5.75, 5.25, 5.25) performance across three different LOs derived from component wise paired comparisons between 10 items. While conferring the same level of transparency for overall rankings as BCJ, this provides a detailed look into the components and how an individual item performs across the LOs. This helps educators identify areas where the candidate may require personalised intervention.

$$H'\left[\pi(i \succ j)\right] = \sum_{d=1}^{D} H\left[\pi_d(i \succ j)\right]. \tag{7}$$

This enables the selection of the most informative pairwise comparison by identifying the pair with the highest total entropy:

$$(i,j) \leftarrow \arg \max_{\forall i,j \in \{1,\ldots,N\} \wedge i \neq j} H'\left[\pi(i \succ j)\right]. \tag{8}$$

Context

Additionally, the dominance of transparency and accountability discourses in UK HE assessment policy has shaped how standards are conceptualised and enacted. However, as Hudson et al. (2017) argue, these discourses often produce a 'conceptual acrobatics' in which assessors feel compelled to speak as though standards are explicit and stable, even when their tacit and contextual nature is widely acknowledged in practice. Their study of external examiners highlights how quality assurance regimes encourage a reliance on codified criteria, while suppressing recognition of the interpretive judgement at the heart of marking. This tension is directly relevant to the uptake of innovative methods such as BCJ and MBCJ, which foreground and systematise CJ as a transparent yet judgement-based approach for assessment. By making visible the role of expert consensus in evaluative decisions, BCJ and MBCJ may help reconcile these competing demands.

These debates are grounded in wider concerns about how assessment standards are established and audited across the UK HE system. Under external examiner regimes and quality frameworks such as the QAA's UK Quality Code for Higher Education, there is pressure to both codify and defend assessment decisions—sometimes at odds with the inherently tacit nature of expert judgement (Strathern, 2000). BCJ and MBCJ challenge this dichotomy by formalising judgement while preserving its comparative, interpretive character. By

surfacing rank probabilities and uncertainty measures, they offer an evidential bridge between professional judgement and institutional accountability.

Thus, in this paper, our goal is to evaluate BCJ and MBCJ in a real-world context with real markers and experts in CJ and see how they fare against absolute marking. We start this discourse with a description of the experimental setup in the next section.

EXPERIMENTAL SETTINGS

In this paper, we consider various factors to determine educators' opinions on conventional grading, standard BCJ, and multi-criteria BCJ. We use coursework submissions that were evaluated formally by the assessment team when the module was delivered.

Participants: Three trained marking assistants with experience in assessing work for this module carried out traditional absolute marking, standard BCJ, and MBCJ. It should be noted that the submissions used in this study were not originally marked by this cohort of assistants; consequently, they encountered all submissions for the first time during the experiment. The full procedure is described in "Research approach" section.

Sample size: The sample size ($N = 10$) was intentionally chosen to reflect the scale of a typical MSc-level cohort within a UK university context, rather than to emulate large-scale or high-stakes assessment settings. While this necessarily limits the generalisability of the findings, it supports a focused and ecologically valid comparison of assessment approaches within a realistic postgraduate teaching scenario. The study is therefore positioned as an exploratory investigation of assessor experience, process and transparency at a scale representative of routine practice, rather than as a definitive evaluation of system-level performance.

Ground-truth scores and ranks: Benchmark scores were derived from the marks originally awarded to students when the assessment was undertaken in a previous academic year. These marks had already been used for summative purposes and released to students, providing a pragmatic and authentic reference grounded in established institutional marking practices. Although this benchmark does not constitute an objective gold standard, it reflects the outcomes of standard moderation and quality assurance processes typically applied in UK higher education and is therefore appropriate for comparative analysis across marking methods.

Time limit: We imposed a time limit on CJ but not on traditional absolute marking. This decision reflects the structural differences between the two approaches. Traditional marking requires a complete set of absolute judgements to produce a full mark profile and therefore could not be meaningfully constrained without undermining the validity of the comparison. In contrast, CJ allows assessors to continue making pairwise judgements indefinitely, albeit with diminishing informational returns. To ensure comparability while maintaining ecological validity, a two-hour stopping point was imposed for CJ. This duration aligns with the common institutional assumption that markers typically spend approximately 10–15 min per script when assessing extended work, meaning that two hours represents a realistic bound for marking a comparable set of submissions. The imposed cap therefore reflects practical marking constraints rather than an arbitrary experimental limitation.

Once the marking assistants had completed all three marking approaches, they were asked to answer a questionnaire (see Appendix A) and, later, are brought together to discuss the approaches used in a workshop (see Appendix B). The techniques used are explained in "Research approach" section. We present the findings to industry experts who carry out research within academia on CJ in an educational setting and people working within industry who look at the policies around assessment for government and exam boards who also research how CJ can be used in educational settings (see Appendix C).

Dataset

We received marks for 30 pieces of work that were submitted at a master's level (UK Level 7 taught-postgraduate) course from the lead lecturer, 'Oracle', where the prompt was to critically review a recent research paper. The submissions were anonymised with any identifying detail (e.g., student ID) removed. We used a distinct identifier with no relationship with the student and a corresponding mark. We created three groups, each with 10 pieces, for traditional rubric-based, BCJ and MBCJ marking, using a stratified sampling approach (Neyman, 1992). The distributions of the marks (out of 20) is given in Figure 5.

The dataset was chosen due to its relevance across educational levels: subjective assessments involving open-ended tasks, such as essays, are common in HE and schools. With a 1000 word limit, these submissions strike a useful middle ground: long enough to allow for depth and complexity, as often required in HE, but not so lengthy as to be unrepresentative of school tasks.

The assessment criteria for the critical review comprised *five* areas, with their relative weights shown in brackets below:

Introduction and Summary (20%): A clear explanation of the topic and a concise summary of the aims, main findings, and key arguments are evaluated.

Quality of Analysis and Evaluation (30%): The assessor looks for originality, a strong evaluation of the paper's strengths and weaknesses, and the identification of any unique aspects.

Conclusions (20%): Assesses the student's ability to summarise the perspectives presented and discuss potential future impact or work.

Writing Quality (20%): Focuses on organisation, structure, coherence, effective transitions, and grammatical accuracy.

References (10%): Examines the student's ability to provide relevant sources and apply correct formatting.

Each criterion is marked out of 100, weighted according to its respective weighting, and combined to produce an overall mark out of 100. This is then converted to a mark out of 20 by rounding to the nearest whole number, which is the version used in this study.

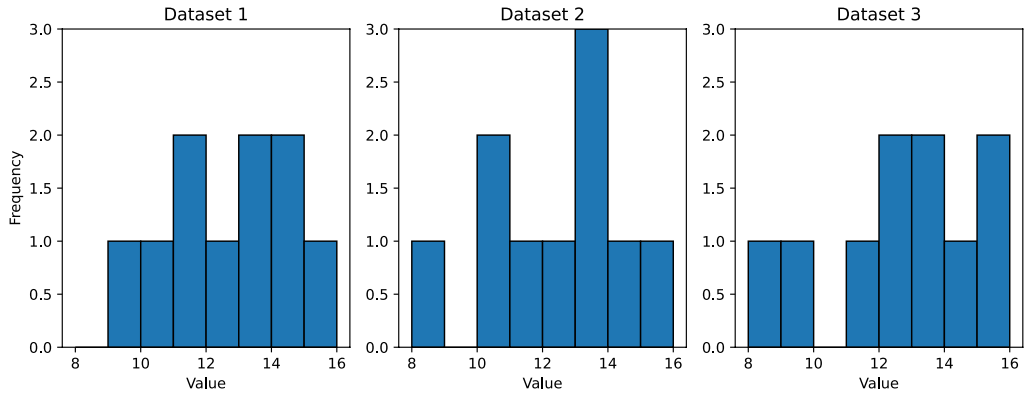

FIGURE 5 Histogram of marks for submissions in different groups: Candidates for traditional marking, BCJ and MBCJ, entitled as Datasets 1, 2 and 3, respectively. Clearly, the groups have similar distribution over the range between 8 and 15; this is important for a fair comparison between the groups. It should be noted that the dataset contains three, four, and three pairs with equal marks in Datasets 1, 2, and 3, respectively, which may make it challenging to recover a fully ordered ranking when only limited comparison data are available.

The assignment brief, which explained the criteria and tasks, was made available to the markers in advance for preparation prior to the experiment.

Web interface for experimentation

BCJ and MBCJ marking were done through web applications. The designs for the applications are similar, with the main difference being the standard BCJ approach only had a single button for each item being displayed. In contrast, the multi-dimension application had a button for each LO that was being compared against, ensuring that only one button could be pressed for one of the items.

In Figure 6, we show the application's interface for the single-dimensional version. The user is presented with two items that are being compared and two buttons. The user presses the button related to the item they deem to be of higher quality. Once the button has been pressed, this updates the result matrix for the entropy-based selection method, and then uses it to select a new pair of items for the assessor to make a judgement on. Assessors can continue until they want to stop. However, it is recommended that a minimum of the number of items (N) times 10 comparisons are completed (Jones & Davies, 2024).

When assessors want to view the ranking of the items, they can view the results page, as demonstrated in Figure 7. The items are rendered in rank order, so the highest ranked item appears first, and the weakest item, as they score it, will be at the bottom of the page. A graph is shown alongside the items depicting the ranking distributions of the items that have been compared. The ranks are calculated from the performance matrix using either BCJ or MBCJ while navigating to the results page.

Figure 8 shows the comparison screen that the assessors view when making their pairwise comparisons. This screen is similar to the standard approach (see Figure 6), but has

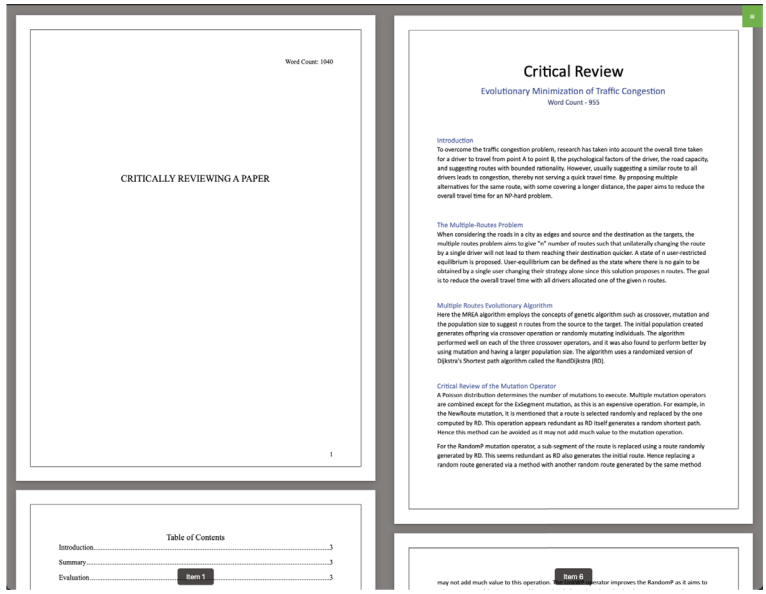

FIGURE 6 An example of the web app page for the standard BCJ's comparison. This page is what the assessor will see when they are making their judgements on the items being presented to them. Once they have pressed the corresponding button linked to the item they prefer, this will update the scores and then produce two new items for the assessor to compare.

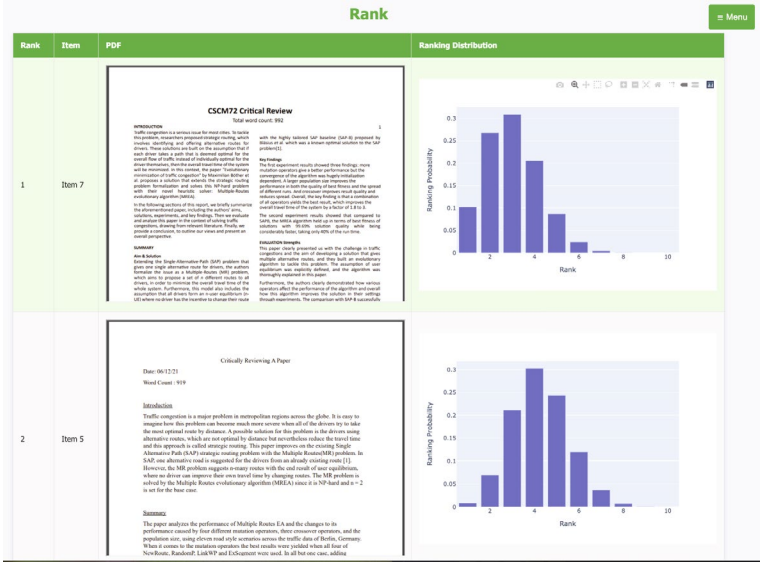

FIGURE 7 When the assessor wants to view the results, they can visit the results page. This web app page shows the items in order of their ranking, so the item ranked first will appear first on the page, and as the assessor scrolls down the page, they will then view the additional items until they reach the last ranking item. Each item's rank, a copy of the item and their ranking distribution are shown to the assessor, ensuring maximum transparency is present to them on how the decisions have been made. This shows the results probability distribution and the ranking after the pairwise comparisons for BCJ.

some key differences. The items have a button for each LO that is being assessed, as well as a submit button. When an assessor presses, for example, the LO1 button for item A, it will light up to show it has been selected; if they were to press LO1 for item 2, the button will dim back to the default colour to ensure that only one LO for each item is selected. Once the assessor is happy with the selections, they hit the submit button and each LO's preference matrices are updated, which enables the differential beta entropy to select the next pair of items to present to the assessor.

Figure 9 presents the results of a multi-criteria Bayesian CJ web app, showcasing transparency in ranking distributions across various LOs. The visualisation provides an overview of the item's overall rank and distribution, and its performance within each LO.

The results section shows the ranking distributions, which are multiple bar charts illustrating the frequency distribution of rankings for the item across all LOs (see Figure 10). The overall ranking distribution is initially shown, similar to the standard BCJ web app, but there is an additional button that enables the user to expand or collapse a dropdown area that enables the ranking and distributions of the individual LOs for each item. Three additional bar charts represent the ranking distributions for each LO, allowing for easy comparison and analysis. The overall rank of the item is displayed alongside the item itself, providing a clear indication of its performance in relation to other items. Showing the distribution of rankings for the item across all LOs offers insight into its performance within each outcome as a holistic overall.

The results section of the web app provides a clear and transparent representation of the item's ranking distributions across various LOs. The overall rank and performance metrics offer valuable insights for educators and researchers seeking to understand the item's strengths and weaknesses in relation to other items.

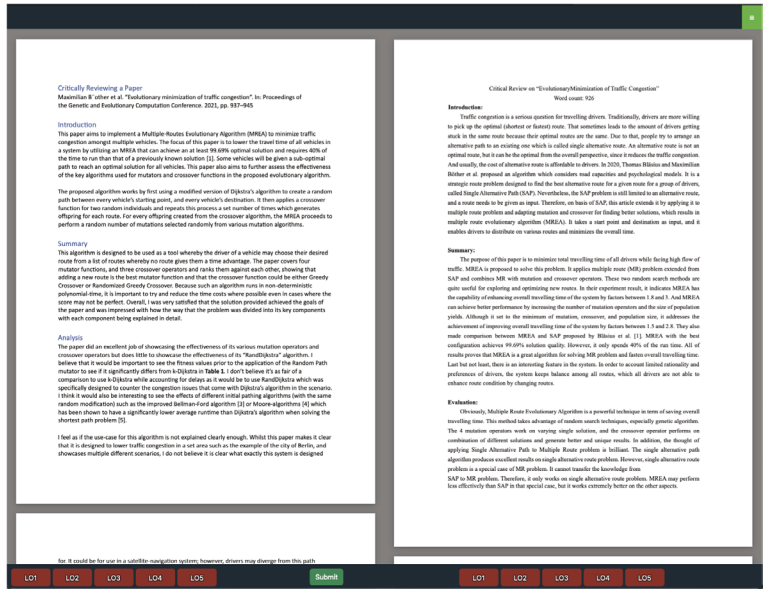

FIGURE 8 Example of the interface for the web app page for the multi-criteria comparison page. Like the standard BCJ page, this is where the assessor will make their decisions on the items displayed. However, they will need to make decisions based on individual LOs this time. They press the submit button once they are ready to submit their preferences. This will update the LOs results and then produce two new items on which to make judgements.

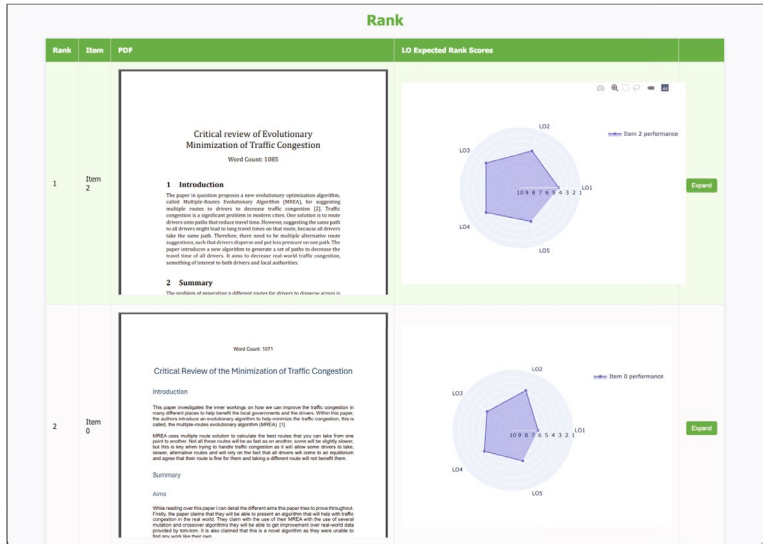

FIGURE 9 Example that presents the results of a multi-criteria BCJ web app showcasing transparency in expected rank E_r scores across different LOs using a radar plot for each item. An expand button is available for the assessor to be able to view the complete rank distributions for the individual LOs.

Research approach

Three markers were recruited for this experiment. The markers were experienced assessors who have been part of the module used in this experiment for a number of years. The markers were given as much time as needed to complete the traditional marking, and a maximum

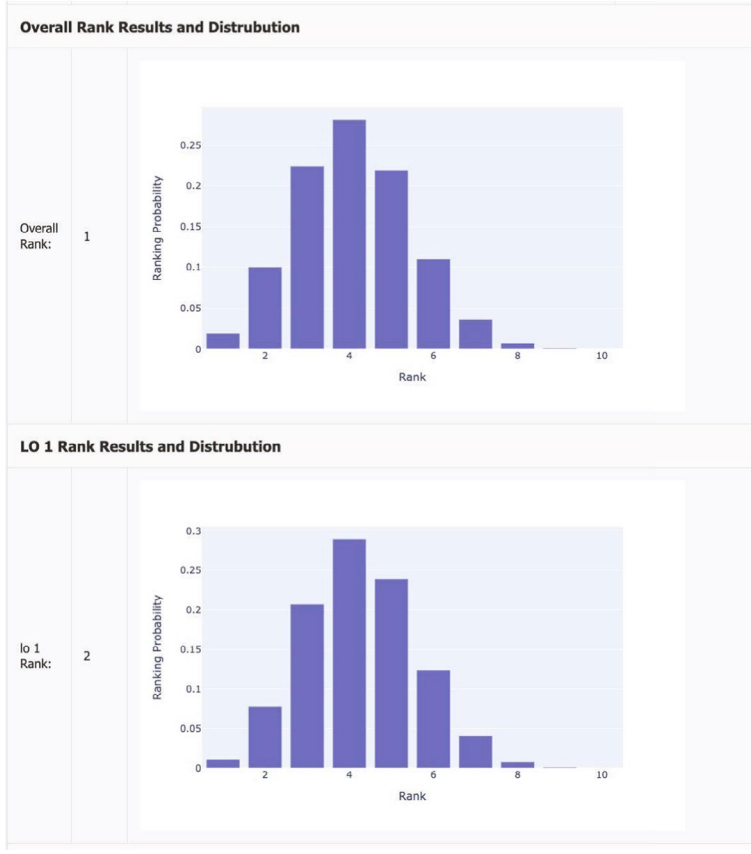

FIGURE 10 Example of the results of a multi-criteria BCJ web app showcasing transparency in ranking distributions across various LOs. The page provides an overview of the item's overall rank and distribution and its performance within each LO.

of 2 h to complete comparisons for each of the CJ methods. The markers did all three methods in different orders to try and mitigate against familiarising with the marking criteria over time. Once the markers had completed them, they took part in the semi-structured questionnaire (see Appendix A) in isolation from the other markers. After this, all three markers came together for an in-person workshop to discuss their experiences as a whole (see Appendix B for an outline of the initial workshop plan).

To further validate and contextualise the findings, three CJ experts were recruited for in-depth interviews. These experts were selected based on their established publication record and practical experience implementing CJ systems in educational settings, with backgrounds from academia, government and industry. Each expert took part in a semi-structured interview, conducted remotely, where they were first presented with an overview of the key results from the marker experiment. All interviews were conducted in isolation to ensure independent feedback, and the findings from these discussions were used to inform the final analysis and discussion (see Appendix C).

Performance metrics

To evaluate the efficacy of a judgement method, two key psychometric properties must be considered: *reliability* and *validity* (Carmines & Zeller, 1979).

Reliability (or consistency) refers to the extent to which variability in observed scores reflects genuine differences rather than measurement error in the context of CJ (Andrich, 2011). As outlined by Gray et al. (2025), reliability may be examined at two levels: *global* and *local*. Global reliability concerns the stability of the overall item ranking, whereas local reliability assesses the consistency with which individual item pairs are judged.

Global reliability can be quantified using *Rank Separation Reliability* (RSR), which is closely related to the well-established *Scale Separation Reliability* (SSR) used in BTM-based CJ. Unlike SSR, RSR operates directly on ranks rather than latent scale scores. Conceptually, RSR measures the extent to which the variability in expected ranks across items exceeds the average model-induced error in their estimation. Formally:

$$\mathrm{RSR}(R) = \frac{\mathrm{Var}\left(\left\{\mathbb{E}[r_i]\right\}_{i=1}^{N}\right) - \frac{1}{N}\sum_{i=1}^{N}\mathrm{Var}(r_i)}{\mathrm{Var}\left(\left\{\mathbb{E}[r_i]\right\}_{i=1}^{N}\right)}, \tag{9}$$

where $R \in \mathbb{R}^{N\times N}$ is the rank-distribution matrix with $R_{i,j} = P(r_i = j)$, and $\mathrm{Var}(\cdot)$ denotes the variance operator.

For local reliability, Gray et al. (2024) introduce two complementary measures of assessor agreement at the pair level: *mode agreement percentage* (MAP) and *expected agreement percentage* (EAP). MAP uses the mode of the Beta posterior for the pairwise preference probability—equivalent to the MLE for a Bernoulli process—and therefore corresponds to the point estimate produced by BTM. It is defined as:

$$\mathrm{MAP}(\alpha_{\mathrm{post}}, \beta_{\mathrm{post}}) = \frac{|\, m(\alpha_{\mathrm{post}}, \beta_{\mathrm{post}}) - 0.5\,|}{0.5} \times 100\%, \tag{10}$$

where the mode is

$$m(\alpha_{\mathrm{post}}, \beta_{\mathrm{post}}) = \frac{\alpha_{\mathrm{post}} - 1}{\alpha_{\mathrm{post}} + \beta_{\mathrm{post}} - 2}.$$

MAP is essentially a reflection of the ratio of wins to total outcomes for a pair, expressed in terms of its distance from complete disagreement at $p = 0.5$. As such, it does not fully capture the remaining uncertainty, which is particularly problematic when only a small number of comparisons have been made for a given pair. Consequently, MAP can fluctuate substantially from one outcome to the next in the early stages.

In contrast, EAP integrates over the full Beta posterior, smoothing across uncertainty and providing a more stable, informative indicator of local reliability. It is defined as:

$$\mathrm{EAP}(\alpha_{\mathrm{post}}, \beta_{\mathrm{post}}) = \kappa \int_0^1 p^{\theta_1}(1-p)^{\theta_2}\,|\,p - 0.5\,|\,dp, \tag{11}$$

where $\kappa = \frac{\Gamma(\alpha_{\mathrm{post}} + \beta_{\mathrm{post}})}{0.5\,\Gamma(\alpha_{\mathrm{post}})\Gamma(\beta_{\mathrm{post}})} \times 100$, $\theta_1 = \alpha_{\mathrm{post}} - 1$, and $\theta_2 = \beta_{\mathrm{post}} - 1$. An analytical closed form of this integral is provided in Gray et al. (2025). Intuitively, EAP quantifies the degree of disagreement and the extent to which the posterior distribution diverges from $p = 0.5$ (the point of complete disagreement). High EAP values indicate convergence towards

consensus on a winning item, whereas values approaching zero signal persistent ambiguity and consistent disagreement.

Validity concerns the extent to which a method accurately recovers the 'true' ordering of items. Gray et al. (2024) focus on *criterion-specific validity*: given a known target ranking, how well can the method reproduce it? In the CJ literature, this form of validity is sometimes referred to as *benchmark reliability* (Crompvoets et al., 2022). Because BCJ and MBCJ produce ordinal outcomes, Kendall's τ is employed—following Gray et al. (2024) and Gray et al. (2025)—as it is the appropriate measure of ordinal agreement and directly evaluates whether the method preserves true pairwise relations. Validity in this form is conceptually aligned with *accuracy*, which cannot be inferred from reliability alone. For example, a weighing scale that is consistently miscalibrated may be highly reliable but inaccurate. Similarly, assessors in CJ may be perfectly consistent while applying an incorrect internal criterion, yielding reliable but invalid rankings. For instance, Leech et al. (2022) demonstrate that assessors were consistent yet incorrect when judging the difficulty of science examination questions.

To quantify this, we compute the normalised Kendall τ distance, which measures discrepancies between two ranked lists. Distances range from 0 (perfect agreement) to 1 (complete disagreement) (Fagin et al., 2003; Kendall, 1938). For example, a distance of 0.03 indicates that only 3% of pairwise orderings differ. In this study, as each method progressed, we recorded the τ distance after every paired comparison, enabling us to examine how quickly each method converged towards the target ranking—serving as an indicator of benchmark reliability or criterion-specific validity. The final τ distance reflects the accuracy, and therefore the validity, of the ranking produced by each methodology.

RESULTS AND DISCUSSION

Before scrutinising the results, it is important to acknowledge that the effectiveness and validity of CJ have already been well established across educational levels—from primary to HE (Bartholomew & Jones, 2021; Jones et al., 2019; Marshall et al., 2020)—particularly in subjects requiring holistic or qualitative evaluation.

BCJ and MBCJ build on this foundation by offering a probabilistic backend that enhances transparency and interpretability without altering the core process of pairwise comparison. As such, many of the proven benefits of CJ in school settings—such as improved reliability, reduced marking time, and increased assessor engagement—are retained in BCJ and MBCJ, with the added advantages of quantifiable uncertainty, clearer insight into ranking rationale, and improved accountability. While sectoral differences in assessment practice exist, the core principles underpinning CJ and BCJ suggest a high degree of transferability, particularly in HE disciplines where traditional rubric-based marking may struggle to ensure fairness and transparency.

Nonetheless, we acknowledge that much of the contextual framing in the literature is grounded in primary and secondary education, whereas our study applies BCJ and MBCJ to a real dataset from HE. It is important to recognise key structural and cultural differences between sectors, such as assessment governance, professional autonomy, and moderation practices. While our findings suggest high transferability, particularly for open-ended tasks assessed via rubrics, further research is needed to explore the implications of BCJ/MBCJ in distinct institutional contexts. We therefore caution against uncritical transferability while noting that many of the underlying principles of fairer, more reliable assessment transcend sector boundaries.

In the remainder of this section, we first examine the performance of individual assessors against the Oracle ground-truth ranks when using the three approaches—traditional absolute marking (TAM), Bayesian comparative judgement (BCJ), and multi-criteria BCJ

(MBCJ)—to evaluate both the comparative effectiveness of each method and the variability in assessors' performance.

We then analyse the assessors' responses to our questionnaires, along with insights gathered from the workshop session held with all assessors and from expert interviews. Our focus is on whether, and in what ways, BCJ and MBCJ support transparency, as well as the issues and considerations raised regarding their use in practice.

Quantitative performance

In Tables 1–3, we present the outputs from the three scoring methods for each marker: traditional absolute marking (TAM), Bayesian CJ (BCJ), and multi-criteria BCJ (MBCJ). The corresponding performance metrics are reported in Table 4, while the relative performance between markers is summarised in Table 5. Throughout, we analyse performance using the convention of reporting the mean accompanied by one standard deviation in the format $a \pm b$, where a denotes the mean and b the standard deviation.

Temporal efficiency

For TAM, markers required on average 112 ± 25 min (1:52 ± 00:25) to complete the task. This aligns with expectations—given that extended-work marking is typically estimated at 10–15 min per item, a total of 100–150 min for 10 items is reasonable. Marker 2 was exceptionally fast, completing the task in just 86 min (1:26).

In contrast, under BCJ, markers required approximately 89 ± 29 min (1:29 ± 00:29) before either reaching the imposed two-hour cap (02:00) or voluntarily stopping once they felt they could no longer make informative comparisons.

For MBCJ, markers required 104 ± 20 min (1:44 ± 00:20), again stopping either at the time limit or due to fatigue. Interestingly, the marker who was the fastest under BCJ was the slowest under MBCJ.

It is notable that each method had a different 'fastest' marker: Marker 2 for TAM, Marker 1 for BCJ, and Marker 3 for MBCJ. Because the methods were experienced in different orders by each marker, this may suggest an influence of sequential exposure or learning effects.

Nonetheless, despite the ordering, BCJ was completed the fastest on average, followed by MBCJ, and finally TAM. The standard deviations were similar across methods, indicating

TABLE 1 Scores by marker with item IDs using dataset 1.

Item	Marker 1	Marker 2	Marker 3
1	11	14	12
2	10	14	13
3	15	14	14
4	13	8	13
5	13	14	16
6	12	12	15
7	13	10	15
8	7	8	12
9	11	10	12
10	16	18	13

TABLE 2 BCJ output for all markers using dataset 2.

Marker	Item	Exp.	SD	P10	P50	P90
1	1	5.69	1.31	4	6	7
	2	7.25	1.30	6	7	9
	3	6.75	1.30	5	7	8
	4	6.75	1.30	5	7	8
	5	5.25	1.30	4	5	7
	6	4.25	1.30	3	4	6
	7	6.88	1.27	5	7	8
	8	3.00	1.24	1	3	5
	9	4.25	1.30	3	4	6
	10	4.94	1.28	3	5	7
2	1	7.75	1.30	6	8	9
	2	6.00	1.32	4	6	8
	3	4.75	1.30	3	5	6
	4	4.75	1.30	3	5	6
	5	7.25	1.30	6	7	9
	6	4.25	1.30	3	4	6
	7	5.00	1.32	3	5	7
	8	3.25	1.30	2	3	5
	9	5.25	1.30	4	5	7
	10	6.75	1.30	5	7	8
3	1	6.12	1.27	5	6	8
	2	7.25	1.30	6	7	9
	3	3.50	1.24	2	3	5
	4	5.75	1.30	4	6	7
	5	5.75	1.30	4	6	7
	6	5.00	1.32	3	5	7
	7	6.75	1.30	5	7	8
	8	3.12	1.27	2	3	5
	9	6.88	1.23	5	7	8
	10	4.88	1.27	3	5	6

Note: We report the expected rank (Exp.) and standard deviation (SD) of each item's rank distribution, together with an 80% high-confidence interval, showing the 10th percentile (P10), median (P50), and 90th percentile (P90) of the posterior rank densities.

broadly comparable variability in marking times. Within the limits of this small sample, BCJ appears to demonstrate the greatest temporal efficiency. Of course, the sample size is insufficient to make strong statistical claims about significance.

Volume of work completed

Under TAM, each marker evaluated all 10 pieces of student work.

Under BCJ, markers completed approximately 48 ± 2 pairwise comparisons—slightly exceeding the total number of unique pairs (45, computed using Equation (1)). Consequently,

TABLE 3 MBCJ output for all markers using dataset 3.

Marker	Item	Exp.	SD	P10	P50	P90
1	1	4.90	1.42	3	5	7
	2	7.40	1.34	6	7	9
	3	4.10	1.39	2	4	6
	4	4.95	1.45	3	5	7
	5	6.35	1.40	5	6	8
	6	4.95	1.44	3	5	7
	7	5.35	1.46	4	5	7
	8	6.00	1.42	4	6	8
	9	5.60	1.45	4	6	7
	10	5.40	1.46	4	5	7
2	1	5.13	1.39	3	5	7
	2	7.03	1.37	5	7	9
	3	3.48	1.31	2	3	5
	4	4.75	1.39	3	5	7
	5	6.96	1.36	5	7	9
	6	5.20	1.46	3	5	7
	7	5.10	1.47	3	5	7
	8	5.92	1.44	4	6	8
	9	5.30	1.48	3	5	7
	10	6.14	1.44	4	6	8
3	1	4.96	1.44	3	5	7
	2	7.70	1.31	6	8	9
	3	4.45	1.43	3	4	6
	4	5.48	1.43	4	5	7
	5	6.69	1.38	5	7	8
	6	5.55	1.43	4	6	7
	7	4.58	1.41	3	5	6
	8	3.95	1.38	2	4	6
	9	5.57	1.45	4	6	7
	10	6.07	1.42	4	6	8

Note: We report the expected rank (Exp.) and standard deviation (SD) of each item's rank distribution, together with an 80% high-confidence interval, showing the 10th percentile (P10), median (P50), and 90th percentile (P90) of the posterior rank densities.

some pairs were necessarily shown more than once. Interestingly, Markers 1 and 3 (not Marker 2, who was the fastest in TAM) completed their BCJ comparisons substantially faster, requiring 35 min fewer or less to reach a similar number of comparisons.

For MBCJ, markers completed approximately 47 ± 16 comparisons. Marker 3 required a similar amount of time for both BCJ and MBCJ, completing three additional comparisons in the latter. Marker 1, however, required nearly an hour longer to complete MBCJ and produced 12 fewer comparisons. In contrast, Marker 2 was more efficient in MBCJ than in BCJ, completing 11 more comparisons in roughly the same amount of time.

Importantly, the two-hour limit ensured that, across markers, nearly every pair received at least one comparison. Our review confirms that all pairs had at least one judgement in BCJ,

TABLE 4 Summary of marker performance under the three scoring methods: Traditional absolute marking (TAM), Bayesian CJ (BCJ) and multi-criteria BCJ (MBCJ).

	TAM		BCJ						MBCJ			
Marker	Time	τ	Time	#Comp	BTM τ	BCJ τ	SSR	RSR	Time	#Comp	τ	RSR
1	2:17	0.36	1:07	49	0.24	0.16	0.67	0.07	2:00	37	0.13	−1.68
2	1:26	0.4	2:00	46	0.2	0.18	0.67	0.06	1:49	57	0.16	−0.97
3	1:54	0.4	1:22	50	0.18	0.18	0.64	0.05	1:22	53	0.29	−0.77
Combined	5:37	0.49	4:29	145	0.04	0.02	0.58	0.6	5:11	147	0.2	0.35

Note: Time is reported in hours:minutes, and #Comp indicates the number of comparisons. BTM-based results (e.g., BTM τ and Scale Separation Reliability, SSR) are only applicable to the holistic comparisons made during the BCJ experiment. Rank separation reliability (RSR) is available only for BCJ and MBCJ. Combined time is obtained by summing assessors' individual times. The combined TAM τ is computed by averaging markers' scores, whereas for BCJ and MBCJ, combined scores are generated by supplying all comparisons from all markers to the corresponding ranking mechanism.

TABLE 5 Relative performance of the markers under the three scoring methods—traditional absolute marking (TAM), Bayesian CJ (BCJ) and multi-criteria BCJ (MBCJ).

		TAM		BCJ			MBCJ		
Marker	Combined	|Δ Time|	|Δτ|	|Δ Time|	|Δτ|	|Δ RSR|	|Δ Time|	|Δτ|	|Δ RSR|
1	2	00:51	0.4	00:53	0.33	0.01	00:11	0.11	0.72
1	3	00:23	0.4	00:15	0.29	0.02	00:38	0.24	0.92
2	3	00:28	0.49	00:38	0.27	0.01	00:27	0.18	0.2

Note: The absolute differences are shown to illustrate how markers varied in time taken and in the quality of the output ranks.

and nearly all pairs did so in MBCJ (with one marker completing eight fewer comparisons than the full set).

If we assume that marking all items under TAM is equivalent to making a decision about every possible pairwise comparison, then the markers effectively achieved full pairwise coverage in BCJ and nearly full coverage in MBCJ. Combined with the earlier temporal analysis, this suggests that BCJ offers greater consistency in the volume of work completed and provides a clear temporal advantage in terms of informational gain per unit of assessor time.

Comparison with ground truth

We first examine the extent to which individual markers—and the aggregated results—agree with the official marks (hereafter the *Oracle*) that were returned to students. While there is no guarantee that this target ground-truth ranking is itself perfectly *valid*, it nevertheless represents the established gold standard within UK higher education for this assessment.

Under TAM, we compared each marker's raw scores with the Oracle using Pearson's correlation coefficient. The correlations were extremely low (0.08, −0.08, and 0.07), which initially appears concerning. However, Pearson correlation is not well suited to ordinal accuracy, so we also computed the Kendall τ distance after converting scores to ranks. As shown in Table 4, the markers' τ distances were 0.36, 0.40, and 0.40. These correspond to 36%, 40%, and 40% of pairwise orderings being discordant relative to the Oracle. Put differently, the markers correctly identified 64%, 60%, and 60% of the pairwise orderings—substantially more reassuring than the Pearson results suggest.

Markers also disagreed considerably with each other: pairwise comparisons showed at least 40% disagreement (i.e., roughly 18 out of 45 pairs). Surprisingly, when we averaged the TAM scores across markers and re-ranked the submissions, the resulting τ distance increased to 0.49—meaning that approximately half of all pairs were incorrectly ordered. Thus, while TAM is considered the gold-standard methodology, these results highlight that subjective assessment of extended written work can vary substantially between assessors even under traditional marking.

Under BCJ, the markers achieved much stronger agreement with the Oracle, with τ distances of 0.16, 0.18, and 0.18. This nearly halves the proportion of discordant pairs relative to TAM: for example, 0.18 corresponds to only about 8 misordered pairs out of 45. When applying the standard BTM (rather than BCJ) to the same comparison data for each marker, the BTM-based τ distances rose (except for Marker 3, where they were tied), changing from 7 to 11 discordant pairs for Marker 1 and from 8 to 10 for Marker 2. This clearly indicates that, given the same data, BCJ produces more accurate rankings than BTM.

When comparisons from all markers were pooled and analysed jointly, performance improved dramatically: the aggregate τ distance fell to 0.02 under BCJ and 0.04 under BTM. In absolute terms, BCJ misordered one pair and BTM misordered two pairs. The difference is small, but BCJ remains marginally superior. More importantly, this demonstrates that CJ performs extremely well when sufficient comparison data are available. It is also striking that markers, who disagreed markedly under TAM, produced aggregate CJ outcomes that aligned almost perfectly with the Oracle.

This alignment should also be interpreted in the context of inter-marker agreement under BCJ: as reported in Table 5, pairwise τ distances between markers decrease by approximately 10% points (4–5 pairs out of 45) relative to TAM. This suggests that BCJ not only improves alignment with the Oracle but also reduces divergence between assessors.

Turning to SSR and RSR, SSR values for individual markers under BCJ are relatively high and close to the recommended threshold of 0.7 (see Kinnear et al. 2025 for discussion). However, SSR drops considerably when data are aggregated. This is surprising given the substantial improvement in ranking accuracy, and raises questions about whether SSR reliably reflects reductions in uncertainty in this context. By contrast, RSR values for individual markers are low—reflecting the wide uncertainty in the rank distributions—but increase meaningfully when aggregated, correctly capturing the reduction in variance of item-wise rank distributions. This suggests that RSR may be a more faithful measure of reliability within a probabilistic ranking framework such as BCJ.

It is also noteworthy that SSR and RSR remain broadly similar across markers when treated independently. This indicates that these metrics primarily reflect the amount of available evidence rather than the accuracy of the derived orderings, reinforcing the argument that validity measures (such as τ distance) are essential alongside reliability metrics, when available.

For MBCJ, Markers 1 and 2 improved their τ distances, each correctly ordering an additional one to two pairs. In contrast, Marker 3 performed considerably worse, misordering approximately five additional pairs.

Interestingly, inter-marker agreement improved within the MBCJ cohort: Markers 1 and 3 disagreed on only 11 pairs which—although the weakest level of agreement within the MBCJ group—still represents better alignment than any pairwise agreement observed under TAM or BCJ. This suggests that explicit criteria may help assessors remain more consistent with one another. However, whether these rankings are *valid* is more complex. Notably, the aggregated MBCJ outcomes showed a substantially higher τ distance (nine misordered pairs) than BCJ (one misordered pair), indicating weaker overall alignment with the Oracle under MBCJ. One possible explanation is the weaker performance of Marker 3 under MBCJ, who also completed their task comparatively quickly.

Finally, RSR values for individual markers under MBCJ were negative. This arises because the variance of the item-wise rank variances exceeded the variance of the item-wise rank means, indicating that uncertainty was too high to confidently distinguish between items. Although the τ distances suggest that mean ranks were still reasonable, the high spread of individual rank distributions reflects genuine ambiguity in the pairwise comparisons. When aggregated, RSR improved substantially (to 0.35), though it remained modest; meanwhile, the aggregate τ distance remained relatively poor.

Overall, BCJ performed the best in recovering the ground-truth ranking: it was the fastest method overall and misordered only a single pair. MBCJ performed less well, misordering nine pairs, although it offers richer insight through criterion-specific ranks and associated uncertainty distributions. TAM performed the worst, misordering approximately half of all pairs, making it the least reliable.

Limitations

We acknowledge several limitations in this study. Although we attempted to create representative sampling across TAM, BCJ, and MBCJ, qualitative differences in the scripts allocated to each group may have influenced the observed outcomes. Furthermore, the experience of the markers differs markedly from that of the Oracle: a substantial proportion of the Oracle's marks were produced by the module leader, who designed the assignment and possesses considerably more marking expertise than the markers in this study. This discrepancy may partly explain the divergence between marker-level and Oracle-level rankings.

In addition, both the dataset and the number of markers were small—albeit broadly representative of a typical taught Master's cohort—which necessarily limits the generalisability of the findings. Nonetheless, these initial results indicate that BCJ and MBCJ hold significant promise as viable approaches to assessment, and they provide a compelling rationale for larger-scale studies involving more students and a broader pool of assessors.

Questionnaire results and analysis

Marker one rated traditional marking as moderately easy to use (three), noting that well-defined criteria made it manageable but still cognitively demanding. Transparency was also rated a three, as the feedback process was clear at an individual level but lacked comparability between students. They were less confident in the accuracy of their marks (two), recognising the potential for inconsistency due to subjectivity and fatigue over time. They approached marking by evaluating each LO separately, weighting them accordingly, but did not particularly enjoy the process. A structured template for students was suggested as an improvement to streamline marking.

BCJ was found to be more difficult than traditional marking, with them rating ease of use as two. They struggled with the holistic nature of the comparisons, as they typically assessed work LO by LO rather than as a whole. Transparency was also rated low (two), as the process lacked clear justifications for the rankings beyond the final distribution. Their confidence in the rankings was rated a three, as comparative ranking helped highlight relative quality but increased subjectivity. They found BCJ more cognitively demanding, especially early on, and would not recommend it over traditional marking.

MBCJ was rated higher in ease of use (four), as LO-delineated comparisons aligned better with their marking approach. They found it significantly more transparent (four), appreciating the radar plot that visualised individual strengths and weaknesses. Their confidence in the rankings was also rated four, as the structured approach reduced subjectivity. Although initially cognitively demanding, they found it became easier over time while maintaining objectivity. They suggested adding an 'unsure' option for cases where two submissions were indistinguishable.

Marker one preferred MBCJ over other methods, as it aligned with how they assessed student work and provided clearer comparative insights. While traditional marking was familiar and felt 'safe', they believed MBCJ had the potential to improve consistency, particularly when multiple markers were involved. They were at their most confident in MBCJ's rankings, as its structured approach reduced inconsistencies in subjective judgement. However, they noted that traditional marking still offered more direct feedback to students, which they felt could be integrated into MBCJ in the future.

Marker two found that traditional marking was the easiest and most transparent method, rating both aspects a five. They appreciated the structured nature of the process, which allowed for clear criteria-based assessment. They noted that providing feedback enhances transparency but mentioned that an even more detailed mark scheme

would be beneficial. However, they were somewhat uncertain about the accuracy of their marks, rating that between three and four. They expressed a preference for structured marking but acknowledged that issues such as inconsistent student presentation could impact the experience.

For BCJ, the participant rated its ease of use a four, citing challenges in comparing papers of similar quality without standard criteria. Initially, they found the method to be exhausting, particularly since it was the first they attempted. Transparency was rated between four and five, as they appreciated the probability distributions but felt it remained somewhat 'black boxy'. They were fairly confident in the ranking results, but noted that their lack of understanding of the underlying algorithm reduced their confidence slightly. They preferred marking individual sections explicitly rather than making holistic judgements.

Regarding MBCJ, marker two found it significantly easier than BCJ, rating it a five for ease of use. They appreciated the ability to compare work across LOs, which made it more transparent than BCJ. Confidence in the rankings was also rated highly, as they could see how individual components contributed to overall scores. However, they pointed out that the method lacked explicit feedback, which they viewed as essential for student improvement.

When asked about their preferred method, they acknowledged that MBCJ was more efficient but favoured traditional marking due to its transparency and ability to provide feedback. They believed BCJ and MBCJ were useful but would work best alongside traditional marking rather than replacing it. Ultimately, they had the most confidence in the rankings generated by traditional marking, as it provided clear, section-by-section scores rather than relative comparisons between students.

Marker three found traditional marking to be the most transparent but also the most time-intensive and mentally demanding. They rated its ease of use as a two, citing the need to apply specific criteria, distinguish between similar scores, and provide feedback. However, they rated transparency as a five, as traditional marking clearly breaks down the reasoning behind each score, though they acknowledged that consistency among markers is crucial. They felt the process was somewhat accurate (three to four) but prone to variability based on the marker's mood or level of fatigue. While they found the approach familiar and structured, they did not enjoy it due to its time-consuming nature and the need to create extensive comments for student feedback.

For BCJ, the participant found it significantly easier, rating it a four or five. They appreciated the simplicity of pairwise comparisons, particularly when differences between submissions were clear. However, they found transparency lacking (two to three), as it was difficult to pinpoint why a particular ranking emerged, especially over time. While seeing the rank distributions helped somewhat, they felt it would not provide enough actionable feedback for students. They were fairly confident (four) in the final rankings, as they aligned with their expectations, though they recognised that inconsistencies in marking could influence results. Compared to traditional marking, they found BCJ generally less mentally taxing, except when comparing closely matched submissions.

Regarding MBCJ, the participant found it more balanced, rating ease of use between three and four. They liked its ability to break down performance across multiple LOs, which made transparency stronger (rated four). They felt this approach provided clearer insights into strengths and weaknesses across criteria. Confidence in the rankings was also high (four), as the method captured differences in individual components while maintaining overall consistency. However, they noted that minor variations, such as differences in referencing, could lead to occasional inconsistencies.

When asked about preferences, the participant found BCJ to be the most straightforward but preferred MBCJ for its ability to highlight strengths and weaknesses across LOs. They believed MBCJ was a strong alternative to traditional marking, especially when multiple markers were involved, as it could help moderate inconsistencies. Ultimately, they had the

most confidence in either traditional marking or MBCJ, with traditional marking being the safer, more familiar option but MBCJ offering potential advantages in efficiency and fairness. They suggested adding a flagging system to indicate particularly difficult comparisons or clear differences to refine the process further.

When looking at all the markers' responses from the questionnaire, we found that traditional marking was associated with high levels of trust and transparency. However, MBCJ was perceived as offering greater transparency than BCJ, primarily because marking was conducted according to each LO. This allowed markers to clearly understand how judgements were made at a granular level, reinforcing their confidence in the method.

MBCJ was generally preferred over BCJ, as markers felt it provided greater insight into the decision-making process. The structure of MBCJ aligned more closely with their usual marking practices, making it a more intuitive approach compared to BCJ. In contrast, BCJ was sometimes perceived as cognitively demanding, particularly when markers encountered two responses they judged to be of equal quality but lacked the ability to flag them as such. This forced them to engage in deeper reflection to make a final decision. Despite this, BCJ was still considered significantly less demanding than traditional marking and only marginally less so than MBCJ. Importantly, the slight increase in cognitive effort required for MBCJ was seen as a worthwhile trade-off, given its perceived transparency. Nevertheless, traditional marking remained the method in which markers placed the greatest trust, particularly regarding final marks and rankings.

Both traditional marking and MBCJ were deemed more transparent than BCJ due to the ability to see how marks were assigned to individual LOs. Markers noted that if traditional marking required only an overall score rather than LO-based marking, its transparency would decrease, making the BCJ approach comparatively more acceptable. This highlights the significance of explicit marking criteria in fostering perceptions of fairness and clarity.

Markers also acknowledged that the comparative nature of BCJ and MBCJ helped mitigate potential biases. In traditional marking, there is a risk that a marker may be overly harsh or lenient in their initial assessments before adjusting their expectations after encountering more responses. The CJ methods counteracted this by requiring markers to make direct comparisons between two pieces of work at a time, thereby reducing inconsistencies arising from fluctuating standards over the marking process.

Across all three interviews, participants generally found traditional marking to be the most transparent but also the most time-consuming and cognitively demanding. While they rated its transparency highly, due to the structured nature of criteria-based assessment, they were less confident in its accuracy, citing concerns about subjectivity, inconsistency, and fatigue over time. They appreciated the ability to provide direct feedback to students but found the process mentally exhausting. Suggested improvements included providing students with structured templates to make marking more efficient and reduce ambiguity.

BCJ was perceived as easier in some respects but introduced new challenges. While some found it straightforward when comparing submissions with clear quality differences, others struggled with its holistic nature, as it did not align with their typical LO-based marking approach. Transparency was rated lower than traditional marking, as the ranking process felt more like a ‘black box’ with limited justification for individual scores. Confidence in rankings varied, with some finding them reasonable but others feeling that the method increased subjectivity. Participants also found BCJ more mentally demanding than expected, especially when comparing closely matched submissions. One participant suggested incorporating an ‘unsure’ option for cases where no clear distinction could be made between two pieces of work.

MBCJ was consistently preferred over standard BCJ and, in some cases, over traditional marking. Participants found it more transparent and easier to use than BCJ, as breaking

down comparisons by LO aligned better with their marking approach. They appreciated the radar plot visualisation, which provided clear insights into students' strengths and weaknesses. Confidence in the rankings was higher than in BCJ, as participants felt that evaluating individual components led to more reliable outcomes. While still cognitively demanding, MBCJ was seen as fairer and more structured. However, they noted that it lacked direct feedback, which they considered essential for students' learning.

Overall, participants preferred MBCJ for ranking work, as it provided more structured comparisons and reduced subjectivity, but traditional marking remained valued for its transparency and feedback. The key takeaway was that MBCJ had strong potential as an alternative assessment method, particularly if mechanisms for providing direct feedback were integrated. Participants also suggested enhancements such as flagging close comparisons, incorporating an 'unsure' option, and using multiple markers to improve consistency.

Workshop results and analysis

The three makers who took part in the experiment came together to discuss their experience as a group while undertaking the marking. At the start of the workshop, when the participants were asked if they felt that the distribution of the samples was evenly distributed between the three sub-samples, they all agreed they were.

The workshop began with a recap of the three marking methods: Traditional Marking, BCJ, and MBCJ. Participants were invited to reflect on their initial assumptions about these methods before reviewing their marking outcomes. Most expected traditional marking to be the most transparent, given its structured, criteria-based approach and the ability to provide direct feedback to students. However, some had concerns about subjectivity and inconsistency, particularly when marking large cohorts. BCJ and MBCJ were seen as less familiar, and there was some scepticism about their fairness and accuracy compared to traditional methods.

The discussion then shifted to participants' experiences with the three marking methods, and their views were consistent with the individual perspectives outlined in the previous section.

After reviewing the ranking outcomes for each method, participants were surprised by the results. Traditional marking had the highest level of inconsistency, with τ scores revealing significant variation between markers. In contrast, BCJ and MBCJ produced rankings that were more consistent and closer to the target rankings. While some had initially believed that traditional marking would be the most accurate, the results suggested otherwise. The relative consistency of BCJ and MBCJ rankings challenged assumptions about the reliability of conventional assessment methods.

The discussion then turned to trust and transparency. Initially, traditional marking was considered the most transparent because it provided explicit scores and justification for each mark. However, after seeing the ranking results, participants questioned whether transparency alone was enough if the method produced inconsistent outcomes. While BCJ and MBCJ lacked direct feedback, they were more reliable in producing fair rankings, which some participants argued could enhance trust in the system. A key challenge remained: how to integrate meaningful feedback into CJ methods.

One of the major concerns was that BCJ and MBCJ, despite their improved consistency, did not provide students with detailed feedback on how to improve. Some participants suggested that automated feedback tools could be developed to provide comments based on ranking decisions. Others proposed a hybrid approach, where BCJ or MBCJ could be used for initial ranking, followed by targeted traditional marking for feedback. This could reduce marking burden while maintaining transparency and student guidance.

Participants also reflected on how marking scales over larger cohorts. Traditional marking was seen as impractical for large groups, as it required significant time and effort to maintain consistency across multiple markers. They discussed how CJ could help mitigate marker bias and inconsistency, particularly if multiple assessors were involved in ranking submissions. MBCJ was seen as particularly useful for moderation purposes, as it allowed different markers to contribute to a more reliable overall ranking. It was perceived that both BCJ and MBCJ would be most effective if it was that multiple markers were working on a larger pool of assessments together, believing that inconsistencies would then be corrected by the BCJ system's ranking abilities. This is an interesting point, as in usual CJ implementations, this is how CJ is usually carried out, as it can be done with one or multiple markers contributing together (Gray et al., 2022). However, this approach was not implemented in this study.

By the end of the workshop, participants had significantly revised their views. Initially, most had assumed that traditional marking was the most trustworthy and accurate method, but the ranking results demonstrated that MBCJ was more consistent and less prone to bias. While BCJ was still viewed as somewhat subjective, MBCJ's structured, multi-criteria comparisons made it a strong alternative to traditional marking. The main limitation remained the lack of direct feedback, which participants felt must be addressed before it could fully replace conventional methods.

The workshop concluded with a discussion on future improvements. Participants suggested that flagging difficult comparisons, incorporating an 'unsure' option, and integrating structured feedback tools could make CJ more effective. They agreed that while traditional marking may remain necessary for providing feedback, MBCJ offered a more scalable, fair and reliable method for ranking student work. The key takeaway was that MBCJ had the potential to replace traditional marking in many contexts, provided feedback mechanisms were developed.

Expert interviews results and discussions

Three experts who research the CJ approach within assessment were interviewed for this section. Two of the experts interviewed work within government educational institutions, with one having previously worked for a UK exam awarding body while researching and implementing CJ; the third was an academic who researches CJ while also implementing it within their teaching practice. The experts were asked questions, in one-to-one semi-structured interviews (see Appendix C).

Expert One (E1) discussed their current use of CJ, primarily for setting grade boundaries rather than direct marking. They noted that CJ is valuable for maintaining transparency and consistency because it focuses expert judgements on comparative quality rather than absolute scores. However, they expressed caution about fully replacing traditional marking, as CJ's holistic nature introduces new biases, such as influences from handwriting or skipped questions, which do not affect absolute marking. Their organisation uses in-house tools for CJ experiments, allowing precise control over the exposure of the person doing the marking to submissions. They also employ rank ordering for efficiency, though it sacrifices some intuitive usability.

Expert Two (E2) actively incorporates CJ into both research and teaching, with these applications sometimes overlapping. Their research focuses on CJ's use in mathematics education, particularly to evaluate problem solving and conceptual understanding. Introduced to CJ in 2009, they have since expanded their work to comparing exam standards and exploring its use in Philosophy, English Literature and Psychology. In teaching, they have used CJ for peer assessment, particularly with undergraduate and foundation mathematics students, and their

early research in 2014 investigated its use in calculus peer assessment. They noted that CJ is engaging and practical for peer assessment, reducing the need to recruit external judges.

Expert Three (E3) provided a detailed account of their experience with CJ, highlighting its evolution and key challenges. While they no longer use CJ extensively in their current role, their early work focused on standard maintenance rather than using CJ as an alternative to traditional marking. Their initial experiences involved manually comparing physical scripts to link standards between different exams, such as A-level Maths syllabi from different decades. This process was slow and logistically complex, prompting them to explore ranking multiple items instead of making only pairwise comparisons. However, they noted that analysis still required converting rankings back into pairs, which could sometimes create a misleading impression of reliability.

Transparency and reliability of CJ

Across the three interviews, there was broad agreement that CJ offers strong *reliability*, though the reasons behind this and its limitations were interpreted differently. Both E1 and E2 emphasised that CJ's reliability arises not simply from its comparative format but from the accumulation of multiple judgements, which helps to minimise individual subjectivity. However, E1 was cautious about overstating CJ's advantages, arguing that when multiple assessors are involved, in traditional marking, its reliability is comparable. Similarly, E2 noted that while some judges (students in particular) have slightly lower reliability, this can be effectively offset by increasing the number of judgements made. E3's reflections were more critical. While acknowledging CJ's strengths, they were concerned about the use of adaptive models, where not all scripts are judged against a common set of comparisons. This, along with potential issues such as expert bias and inadequate attention to differences in test difficulty, was seen as a potential threat to the overall fairness and reliability of the process.

On the question of *transparency*, views diverged strongly. In discussing transparency challenges, E1 pointed out that the holistic nature of the CJ inherently reduces transparency because judges make comparative decisions without justifying their reasoning. E2 saw CJ as neither more nor less transparent than traditional assessments, though they recognised that the lack of detailed mark breakdowns, such as those found in rubric-based systems, can lead to perceptions of opacity. However, they suggested that when CJ is carefully embedded into the learning process, students generally accept it without issue. In contrast, E3 described CJ as often functioning like a ‘black box’, particularly in how relative judgements are translated into final scores. They stressed the importance of clear communication and narrative-building to support the credibility of CJ outcomes, especially for wider audiences unfamiliar with the method. E1 also noted transparency concerns from educators who worry about inconsistent application of criteria between judges. They questioned the value of CJ in contexts where only a single judge is involved, suggesting that such use undermines both reliability and perceived fairness.

Initial views on BCJ and MBCJ

All three experts acknowledged that BCJ builds on standard CJ through a more sophisticated statistical model, but there was also consensus that it does not inherently improve transparency. Both E1 and E2 independently noted that the Bayesian nature of BCJ is largely inaccessible to most users, particularly when they are unaware of how rankings are estimated or how the model handles uncertainty. E1 argued that BCJ remains as opaque as

other statistical methods in CJ and saw little transparency gain unless users are trained in or informed about the algorithmic processes involved. E2 held a similar view, predicting that BCJ would still be perceived as opaque, particularly due to the absence of explicit marking criteria and a clear audit trail.

E1 also raised an ethical dimension, warning that the use of priors in BCJ could introduce bias in high-stakes settings, though they were more accepting of their use in formative assessments, where fairness is less critical and efficiency more important. Their overall conclusion was that BCJ does not enhance transparency or decrease it but does offer a more refined estimation process, which could have practical value depending on the context.

E3 showed interest in BCJ's methodological potential but posed technical questions about how distributions and prior knowledge are modelled over time. While not outright critical, their reflections implied uncertainty about how understandable or explainable BCJ outputs would be without substantial training. They highlighted that while the visual outputs of BCJ might aid interpretation, their full value depends on the user's ability to grasp the underlying logic. Like the others, E3 flagged scalability concerns, questioning the feasibility of BCJ in large-scale assessment contexts like national exams, despite recognising its success in small-scale university settings.

All three experts viewed MBCJ as a promising development, particularly in relation to transparency and alignment with established educational practices. E1, E2, and E3 each highlighted how breaking down judgements by LO makes the process feel more familiar and intuitive to educators. This multi-criteria structure was seen as a strength, enabling judges to evaluate distinct dimensions of quality, such as structure, argumentation or engagement, more explicitly than in standard CJ or BCJ. E1 and E3 both praised MBCJ for enhancing transparency, with E1 stating that the clearer structure reduces holistic subjectivity and makes it easier for judges to articulate their reasoning. E3 similarly noted that MBCJ more closely mirrors traditional assessment logic, where individual attributes of a submission are considered independently. E2 agreed that MBCJ could make the decision-making process more transparent, but they remained unsure whether it fully addresses the lack of an audit trail.

In terms of usability and marker experience, E1 referenced markers' feedback showing that MBCJ was preferred over BCJ due to reduced cognitive load and clearer decision-making, particularly when assessing close calls. The model was seen as easier to use and more natural for those accustomed to rubric-based marking. E2 shared enthusiasm for MBCJ's potential in structured exam settings, echoing E1's view that its design better supports educational assessment. E3 also saw value in its potential to mitigate snap judgements by encouraging assessors to consider each criterion in turn.

However, E2 expressed reservations about interdependencies between criteria, arguing that for research purposes, it is preferable to maintain independence across dimensions. They emphasised the importance of context-sensitive assessment, suggesting that MBCJ should be considered as one tool among several, rather than a universal solution. E3 raised concerns about scalability, questioning how MBCJ would function in large-scale assessment environments like national exams. They noted that while visualisations produced by MBCJ were useful, significant training would be needed for assessors and stakeholders to fully understand and trust the outputs.

In comparison to CJ and BCJ, all three experts agreed that MBCJ retains the core advantages of reliability and efficiency but adds improved transparency and greater alignment with traditional practice. E1 viewed this as a way to build trust with educators, especially if MBCJ's outputs can be paired with familiar statistical metrics. E2 and E3 both saw it as a step forward in design, even if implementation at scale and full transparency remain unresolved challenges.

Future directions

A key theme was the shift in perception after reviewing BCJ and MBCJ results. Initially, traditional marking was preferred for its perceived transparency, while BCJ was seen as reliable but opaque. However, MBCJ emerged as the preferred approach, maintaining high reliability while reducing cognitive strain and providing clearer decision-making structures. Experts noted that this change in preference underscored the importance of usability and training in the success of new assessment models. Across the three interviews, there was a shared view that MBCJ offers a strong foundation for future development, especially in educational contexts, but that several challenges must be addressed for it to be widely adopted and effectively implemented. There was consensus that MBCJ is particularly well suited to educational contexts, though its role in research was more contested.

However, all three experts identified feedback and usability as priority areas for further work. E1 and E3 both emphasised the need to improve feedback mechanisms within CJ frameworks. E1 saw detailed, criterion-level feedback as a natural extension of MBCJ, aligning with its multidimensional structure, while E3 flagged the challenge of providing meaningful feedback more broadly in CJ-based models. E1 suggested any widespread adoption of MBCJ will require clearer visualisations, training, and integration with existing assessment infrastructures.

Tool development and standardisation was a shared concern. E1 advocated for the creation of open-source tools to reduce fragmentation across CJ implementations. They argued that aligning MBCJ metrics with established CJ reliability statistics would not improve uptake or make it easier to compare outcomes with traditional methods. E3 expressed similar interest in accessible resources and documentation, requesting access to research papers and calling for better support for practitioners engaging with the models.

In terms of scalability and generalisability, E3 looked ahead to the use of BCJ and MBCJ in large-scale assessments such as national exams, identifying this as a critical test for the methods. E2 found MBCJ promising for structured exam marking, but had reservations about using it for research, due to potential interdependencies between criteria. They advocated for a flexible, pluralistic approach, where CJ, BCJ, and MBCJ are seen as complementary tools, each suited to different contexts and purposes, rather than as a single preferred standard. E3 called for more research into how judges process information and make decisions within these systems, especially when applied at scale.

A key theme was the shift in perception after reviewing BCJ and MBCJ results. Initially, traditional marking was preferred for its perceived transparency, while BCJ was seen as reliable but opaque. However, MBCJ emerged as the preferred approach, maintaining high reliability while reducing cognitive strain and providing clearer decision-making structures. Experts noted that this change in preference underscored the importance of usability and training in the success of new assessment models.

BCJ transparency in the assessment procedure

Initially, traditional marking was perceived as the most transparent because it provided a structured, criteria-based approach with explicit marks and feedback. Participants felt that transparency came from clearly defined assessment rubrics, where each score was justified based on LOs. However, they also acknowledged that traditional marking relied heavily on the individual marker's judgement, which could introduce inconsistencies between assessors. Some participants noted that transparency was undermined by subjectivity, as different markers might interpret criteria differently, particularly in open-ended or qualitative assessments.

BCJ was widely seen as less transparent than traditional marking, as it lacked explicit justifications for ranking decisions. Participants found it difficult to determine why one submission was ranked higher than another, as the process was holistic and comparative rather than criteria-based. The ranking system felt somewhat like a 'black box', where the final order of submissions emerged without a clear rationale for individual placements (see Figure 7), apart from the transparency that the systems produce in displaying the ranking probabilities. This lack of insight made some participants feel less confident in the fairness of BCJ, even though the method produced more consistent rankings than traditional marking.

MBCJ, however, was seen as a step towards greater transparency (see Figures 9 and 10). Since it broke down comparisons across multiple LOs, participants felt that the ranking process was more structured and aligned with how they naturally assessed student work. Unlike BCJ, MBCJ provided clearer insights into why one submission was stronger in specific areas, which made it easier to justify ranking decisions. While MBCJ did not provide direct explanations for individual scores, its structured nature reduced the perception of randomness in the process, making it feel more transparent than BCJ.

A major issue discussed was the role of feedback in transparency. Traditional marking was still preferred in terms of transparency because it allowed markers to explicitly communicate reasoning to students. In contrast, BCJ and MBCJ, despite being more consistent in ranking, did not naturally provide detailed feedback on areas for improvement. Participants felt that without feedback, transparency was limited, as students would not fully understand why they received a particular ranking or how they could improve. This was seen as a critical barrier to adopting CJ methods in student assessments.

Participants agreed that transparency must be balanced with reliability and fairness. While traditional marking was still valued for clear justification and student feedback, its inconsistencies reduced trust in the process. BCJ was viewed as too opaque for individual assessments, but MBCJ provided a reasonable compromise by offering structured rankings across criteria. The consensus was that MBCJ had the potential to be a transparent and fair alternative to traditional marking, but only if mechanisms for providing student feedback were integrated into the process.

By using the BCJ or MBCJ approach, we can also display the transparency in the decisions being made, by identifying instances of inter- or intra-marker discrepancy (see Figure 11). If the markers agree about the outcome, the mode of the distribution will be close to 1 or 0, depending on whether they prefer item a or b. However, if the mode is close to 0.5, this indicates disagreement between the markers—or possibly that a single marker is inconsistent with themselves, with 0.5 indicating that 50% preferred item a, while 50% preferred item b. Ultimately, we believe this offers enhanced transparency and greater detail in the presentation of results.

Gray et al. (2025) proposed new metrics to measures agreements between the markers, namely, the mode agreement percentage (MAP) and the expected agreement percentage (EAP). If the MAP (or EAP) is equal or greater than 0.5, then that means the markers are mostly in agreement and that their decisions are outside of the 25th–75th quartile range that represents the band within which there is a high level of disagreement on showing a preference for one item over the other. This is what we would expect if all markers agree that one item is better than another, as displayed in Figure 11, where $i=0$ and $j=5$ with a MAP of 100% and an EAP of 62.5%. It should be noted that A score of 1 (or 100%) represents perfect agreement between the markers. While any value less than 0.5 indicates that the judgements are within the 25th–75th quartile range, representing disagreements. For instance, in Figure 11, where $i=0$ and $j=6$, disagreements between the markers is evident, with a slight overall preference towards the item i with low agreement scores of an MAP 33.3% and an EAP of 37.5%. We can additionally create a heatmap that produces these scores (see Figure 12) to visually identify the pairs that are dividing the crowd.

FIGURE 11 This shows the transparency in the decisions being made. The closer the distribution is to 0.5 (the black line), the more uncertain the markers are, meaning that half went one way and the other half went the other way. The red dotted line represents the mode β value from the decisions made, while the blue line shows the probability density function of the β distribution.

Expert two, when shown these outputs (Figures 11 and 12), found the insights informative and suggested that these metrics could be a good alternative to measuring reliability compared to the current approach of Scale Separation Reliability (SSR).

In order to make the process not only transparent for the assessors but also the students, we propose that students be granted access to key outputs of the BCJ and MBCJ models, including their position within the final ranking and associated decision distributions. By making visible the ranking distributions associated with each judgement and showing how their work compared to others across specific learning outcomes, students can gain a more meaningful understanding of how decisions were reached. This opens up new possibilities for transparency that go beyond fixed rubrics, allowing students to engage with both the outcome and the process of assessment.

Additionally, we recognise that MBCJ could be integrated with more feedback-rich approaches, such as annotated exemplars or structured narrative comments, which would

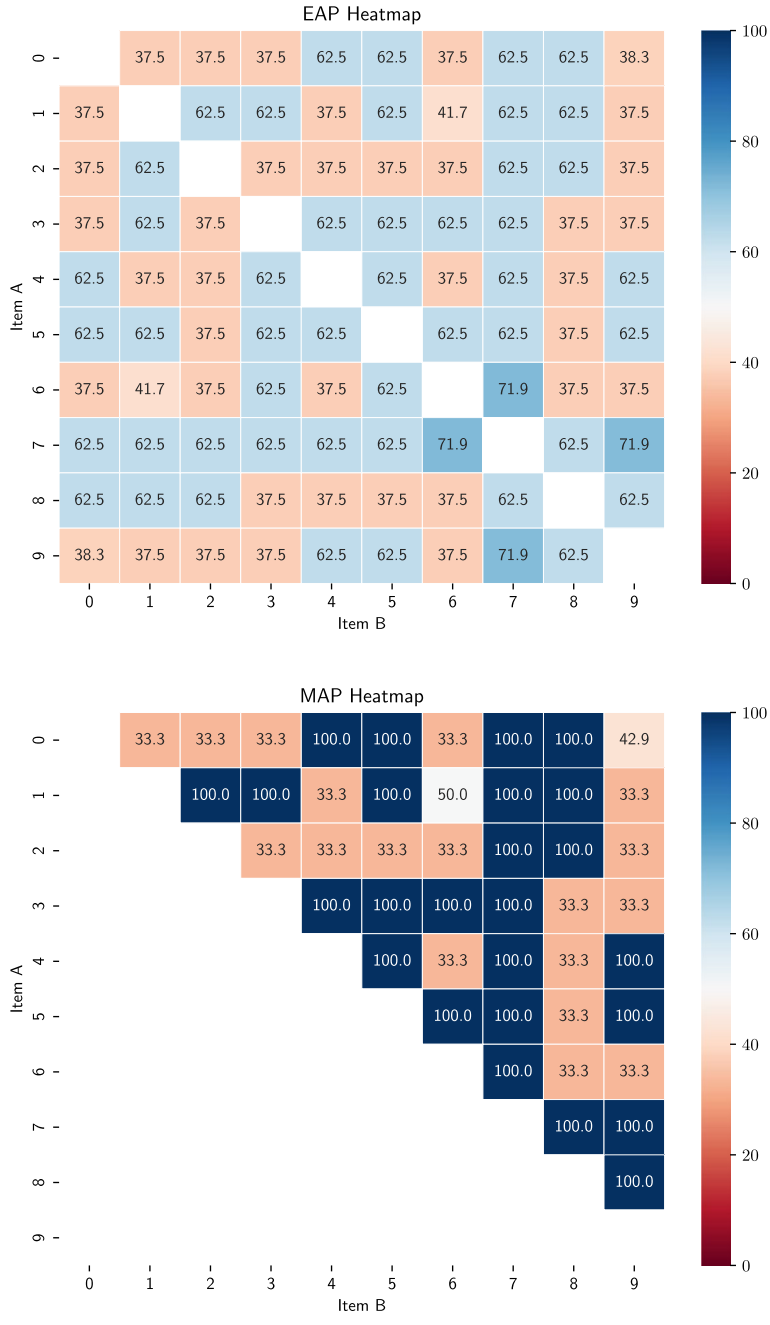

FIGURE 12 This shows an example of the EAP and MAP outputs. These heatmaps can be produced for all LOs for the MBCJ and holistically for the BCJ approach. Any value ≥ to 50 shows that the agreement is outside the 25th and 75th percent quantile ranges.

further support student understanding and actionability. Future work should explore how these combined approaches might improve student trust, clarity, and engagement with assessment.

Beyond its practical utility, the use of MBCJ invites a rethinking of what transparency in assessment means. Traditional notions of transparency often rely on fixed rubrics and

criteria that are assumed to provide clarity and fairness. However, MBCJ challenges this by showing that transparency can also emerge from the structure of decision-making itself through traceable judgement pathways, quantifiable uncertainty, and explicit representations of disagreement. This shifts the focus from merely showing what was assessed to how and why certain rankings emerge, foregrounding the epistemic processes behind evaluation. It prompts us to question whether transparency should be rooted in the visibility of criteria alone or in the interpretability of human judgement within complex domains.

Theoretically, MBCJ reinforces the idea that assessment is not only a technical process but also a form of knowledge construction, for both the assessor and the student. By modelling judgement as probabilistic and data-driven, it disrupts assumptions that expert consensus is static or inherently valid. Instead, it opens space for acknowledging assessor subjectivity and embracing uncertainty as an integral and transparent component of fair assessment practice.

BCJ and MBCJ also speak directly to national concerns around assurance and legitimacy in marking practices. As Bloxham et al. (2016) note, moderation in HE often functions as a ritualised process to satisfy regulatory optics rather than genuinely support standards. The systematic audit trails and uncertainty metrics embedded in BCJ/MBCJ offer a more substantive basis for moderation and review, aligning with contemporary policy demands for transparency while preserving professional discretion.

To summarise, we believe BCJ and MBCJ render greater transparency in the following manner:

- *Pairwise uncertainty made explicit*: Each pairwise decision is represented by a Beta posterior distribution. With finite data, this distribution retains non-zero width, directly reflecting the remaining uncertainty in the preference between two items. These posteriors support transparency in two ways:
  1. they guide the adaptive, entropy-based selection of the next pair to compare and
  2. they highlight pairs where assessors disagree, thereby supporting chief-assessor adjudication when comparisons are difficult or contentious.
- *Interpretable, uncertainty-aware rank outputs*: Unlike standard CJ, which outputs latent scores that may not correspond to interpretable marks, BCJ produces a *predictive rank distribution* for each item, together with explicit uncertainty around that rank. These rank distributions are directly interpretable and can be mapped to letter-grade categories via the principled procedure described in Gray et al. (2024).
- *Criterion-specific transparency through MBCJ*: MBCJ extends BCJ by modelling each criterion separately and then aggregating these criterion-level rank distributions into an overall ranking. This enables assessors and stakeholders to inspect how rankings vary across individual LOs, thereby adding a further layer of transparency to the decision process.
- *Visibility of the consequences of decisions*: Taken together, BCJ and MBCJ provide transparency not by simplifying the underlying computation, but by offering rich, interpretable outputs that make the consequences of assessor decisions visible. Assessors can see where uncertainty remains, where disagreements arise, and how individual comparisons influence the overall ranking outcome.

Implementing BCJ

While the web apps used within this experiment are open-sourced and available on GitHub (see “Conclusions” section for links), the documentation has been provided to make the

process as seamless as possible. However, at the point of writing, there are elements of the web app that the users will have to adapt to use for themselves manually. These changes are explained within the GitHub repository's README file. However, while no great deal of coding knowledge is required, having coding experience will undoubtedly help with the process.

Additionally, for large sample sizes, the ranking process can be resource-heavy. Therefore, depending on the specifications of the machines being used, the ranking process can take some time. Still, the process for both standard and MBCJ for comparing and deciding on the next pair to present to the assessor is relatively quick.

Considering the core elements of BCJ and MBCJ—pair selection, winner determination, and rank generation (as shown in Figure 1)—the pair selection and rank generation steps are performed computationally. When scaling the method to a large number of items, even with limited computational resources, pair selection remains efficient. For example, on a standard desktop machine, selecting the next pair to present to the assessor (based on maximum uncertainty) takes less than 10 milliseconds with 300 items: an amount that could represent a large undergraduate cohort. To put this into context, this is significantly faster than the recommended latency for interactive systems, which is around 100 milliseconds (Attig et al., 2017).

If it is necessary to generate ranks after each comparison, a straightforward Monte Carlo (MC) version of the probabilistic computation for BCJ (Gray et al., 2024) can be used. In this case, with 300 items, the same machine can generate ranks within approximately 20 s. For MBCJ (Gray et al., 2025), this scales linearly with the number of criteria. While this may be too long for real-time interaction—given that acceptable response times for web applications are typically reported to be between 10 and 15 s (Attig et al., 2017)—user experience is highly context-dependent. Further user studies are needed to determine what constitutes a reasonable latency for BCJ and MBCJ in practice.

That said, more efficient computational alternatives to standard MC exist. The simplest among them is quasi-Monte Carlo (QMC), which can improve performance by up to an order of magnitude without significant loss of accuracy (Caflisch, 1998). Future work will explore faster numerical approaches to bring computation times to acceptable levels, alongside user studies to establish reasonable response time expectations in this context.

CONCLUSIONS

Traditional marking is familiar but cognitively demanding and inconsistent. CJ-based methods offer a more structured, consistent, and fairer alternative that reduces subjectivity and aligns well with educators' practices. While BCJ enhances transparency for students and assessors by making ranking distributions visible, MBCJ builds on this by breaking assessments down by LOs, offering greater insight into specific performance areas. MBCJ requires more cognitive effort due to its multidimensional nature, but our participants found the added transparency and clarity worthwhile, though standard BCJ remains a viable option for those seeking a simpler approach.

However, both BCJ and MBCJ lack detailed feedback for students and work is needed to integrate BCJ and MBCJ into large-scale assessments. This includes improving feedback mechanisms, supporting interpretability and exploring how students respond to transparency and uncertainty metrics. MBCJ's structure presents opportunities to address this by generating criterion-specific feedback, especially through automation.

Overall, structured CJ methods—particularly MBCJ—show strong potential to enhance educational assessment by improving transparency, consistency, and workload efficiency, provided they are supported by further development and research.

AUTHOR CONTRIBUTIONS

Andy Gray: Conceptualization; methodology; software; data curation; investigation; validation; formal analysis; visualization; writing – original draft; writing – review and editing. **Alma Rahat:** Conceptualization; supervision; writing – review and editing; project administration. **Stephen Lindsay:** Writing – review and editing; supervision. **Jen Pearson:** Writing – review and editing; supervision. **Tom Crick:** Supervision; writing – review and editing.

ACKNOWLEDGEMENTS

Andy Gray was funded by the EPSRC Centre for Doctoral Training in *Enhancing Human Interactions and Collaborations with Data and Intelligence-Driven Systems* (EP/S021892/1) at Swansea University. Additionally, the project stakeholder is CDSM. We are particularly grateful to their CIO, Darren Wallace. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission. All underlying data to support the conclusions are provided within this paper.

FUNDING INFORMATION

The authors have nothing to report.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The code for the single-dimension app can be found here: https://github.com/codingWithAndy/BayesCJ-Web-App. The code for the multi-criteria app can be found here: https://github.com/codingWithAndy/BayesCJ-multi-dimensional-Web-App.

ETHICS STATEMENT

Ethics approval for the use of the secondary data was obtained from the Faculty of Science and Engineering ethics committee at Swansea University (Research Ethics Approval Number: 12023 7465 6926).

ORCID

Andy Gray https://orcid.org/0000-0002-1150-2052
Alma Rahat https://orcid.org/0000-0002-5023-1371
Stephen Lindsay https://orcid.org/0000-0001-6063-3676
Jen Pearson https://orcid.org/0000-0002-1960-1012
Tom Crick https://orcid.org/0000-0001-5196-9389

REFERENCES

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research*, *11*(5), 571–585.

Attig, C., Rauh, N., Franke, T., & Krems, J. F. (2017). System latency guidelines then and now – Is zero latency really considered necessary? In *Proceedings of International Conference on Engineering Psychology and Cognitive Ergonomics (EPCE 2017)* (pp. 3–14). https://link.springer.com/chapter/10.1007/978-3-319-58475-1_1

Bamber, M. (2015). The impact on stakeholder confidence of increased transparency in the examination assessment process. *Assessment & Evaluation in Higher Education*, *40*, 471–487. https://doi.org/10.1080/02602938.2014.921662

Bartholomew, S., & Jones, M. D. (2021). A systematized review of research with adaptive comparative judgment (ACJ) in higher education. *International Journal of Technology and Design Education*, *32*, 1159–1190. https://doi.org/10.1007/s10798-020-09642-6

Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, *34*, 209–220. https://doi.org/10.1080/02602930801955978
Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK higher education grading practices. *Studies in Higher Education*, *36*, 655–670. https://doi.org/10.1080/03075071003777716
Bloxham, S., Hughes, C., & Adie, L. (2016). What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. *Assessment & Evaluation in Higher Education*, *41*(4), 638–653. https://doi.org/10.1080/02602938.2015.1039932
Bramley, T. (2015). Investigating the reliability of adaptive comparative judgment. Cambridge Assessment Research Report.
Bramley, T., & Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *26*(1), 43–58. https://doi.org/10.1080/0969594X.2017.1418734
Caflisch, R. E. (1998). Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, *7*, 1–49.
Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Sage Publications.
Caron, F., & Doucet, A. (2012). Efficient Bayesian inference for generalized Bradley–Terry models. *Journal of Computational and Graphical Statistics*, *21*(1), 174–196.
Christodoulou, D. (2025). Apw year 5—How did the AI judges do? https://blog.nomoremarking.com/apw-year-5-how-did-the-ai-judges-do-db731dcf84d4
Crick, T. (2021). COVID-19 and digital education: A catalyst for change? *ITNOW*, *63*(1), 16–17. https://doi.org/10.1093/itnow/bwab005
Crompvoets, E. A., Béguin, A. A., & Sijtsma, K. (2022). On the bias and stability of the results of comparative judgment. In *Frontiers in education* (Vol. *6*, 788202). Frontiers Media SA. https://doi.org/10.3389/feduc.2021.788202
de Moira, A. P., Massey, C., Baird, J., & Morrissy, M. (2002). Marking consistency over time. *Research in Education*, *67*, 79–87. https://doi.org/10.7227/RIE.67.8
Department for Education. (2019). Teacher workload survey 2019: Main report. https://assets.publishing.service.gov.uk/media/5e12fcb7e5274a0f9e82e4fd/teacher_workload_survey_2019_main_report_amended.pdf
Department for Education. (2022). School workload reduction toolkit. Retrieved December 29, 2024. https://www.gov.uk/guidance/school-workload-reduction-toolkit#how-to-use-the-toolkit
Department for Education. (2024). Working lives of teachers and leaders: Wave 3 summary report. https://assets.publishing.service.gov.uk/media/674ddf916f6baefc2a9ca1aa/Working_lives_of_teachers_and_leaders_wave_3_summary_report.pdf
Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Lucini, B., Medaglia, R., le Meunier-FitzHugh, K., le Meunier-FitzHugh, L. C., Misra, S., Mogaji, E., Sharma, S. K., Singh, J. B., Raghavan, V., Raman, R., Rana, N. P., Samothrakis, S., Spencer, J., … Williams, M. D. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *53*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002
Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E., Jeyaraj, A., Kar, A., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Ahmad Albashrawi, M., & Balakrishnan, J. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642
Erturk, S., van Tilburg, W. A., & Igou, E. R. (2022). Off the mark: Repetitive marking undermines essay evaluations due to boredom. *Motivation and Emotion*, *46*(2), 264–275.
Fagin, R., Kumar, R., & Sivakumar, D. (2003). Comparing top k lists. *SIAM Journal on Discrete Mathematics*, *17*(1), 134–160. https://doi.org/10.1137/S0895480102412856
Gonsalves, C., & Lin, Z. (2025). Clear in advance to whom? Exploring 'transparency' of assessment practices in UK higher education institution assessment policy. *Studies in Higher Education*, *50*(7), 1454–1470.
Goossens, M., & De Maeyer, S. (2017). How to obtain efficient high reliabilities in assessing texts: Rubrics vs comparative judgement. In *International conference on technology enhanced assessment* (pp. 13–25). Springer. https://doi.org/10.1007/978-3-319-97807-9_2
Gray, A., Rahat, A., Crick, T., & Lindsay, S. (2024). A Bayesian active learning approach to comparative judgement within education assessment. *Computers and Education: Artificial Intelligence*, *6*, 100245. https://doi.org/10.1016/j.caeai.2024.100245
Gray, A., Rahat, A., Crick, T., & Lindsay, S. (2025). Bayesian active learning for comparative judgement: Estimating reliability and managing multiple criteria with applications in educational assessment. https://arxiv.org/abs/2503.00479
Gray, A., Rahat, A. A., Crick, T., Lindsay, S., & Wallace, D. (2022). Using Elo rating as a metric for comparative judgement in educational assessment. In *Proceedings of the 6th International Conference on Education and Multimedia Technology. ICEMT '22* (pp. 272–278). ACM.

Grissom, J. A., Kalogrides, D., & Loeb, S. (2015). Strategic staffing? How performance pressures affect the distribution of teachers within schools and resulting student achievement. *American Educational Research Journal*, *54*, 1079–1116. https://doi.org/10.3102/0002831217716301
Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., & Wang, S. (2024). A survey on LLM-as-a-judge. arXiv Preprint arXiv:2411.15594.
Guskey, T. (2024). Addressing inconsistencies in grading practices. *Phi Delta Kappan*, *105*, 52–57. https://doi.org/10.1177/00317217241251883
Hardman, J., Watermeyer, R., Shankar, K., Suri, V., Crick, T., Knight, C., & Chung, R. (2022). "Does anyone even notice us?" COVID-19's impact on academics' well-being in a developing country. *South African Journal of Higher Education*, *36*(1), 1–19. https://doi.org/10.20853/36-1-4844
Hasan, A., & Jones, B. (2024). Assessing the assessors: Investigating the process of marking essays. *Frontiers in Oral Health*, *5*, 1272692. https://doi.org/10.3389/froh.2024.1272692
Hausdorff, H., & Farr, S. (1965). The effect of grading practices on the marks of gifted sixth grade children. *Journal of Educational Research*, *59*, 169–172. https://doi.org/10.1080/00220671.1965.10883325
Holmes, S., Black, B., & Morin, C. (2020). *Marking reliability studies 2017: Rank ordering versus marking—Which is more reliable?* Ofqual.
Hudson, J., Bloxham, S., den Outer, B., & Price, M. (2017). Conceptual acrobatics: Talking about assessment standards in the transparency era. *Studies in Higher Education*, *42*(7), 1309–1323. https://doi.org/10.1080/03075079.2015.1092130
Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *Annals of Statistics*, *32*(1), 384–406.
Ilahi, F., Manzoor, T., & Elahi, I. (2024). Enhancing assessment integrity: A critical analysis of transparency and fairness in marking process at University of Sargodha. *Journal of Education and Social Studies*, *5*, 489–501. https://doi.org/10.52223/jess.2024.5229
Jerrim, J., & Sims, S. (2021). When is high workload bad for teacher wellbeing? Accounting for the non-linear contribution of specific teaching tasks. *Teaching and Teacher Education*, *105*, 103395. https://doi.org/10.1016/J.TATE.2021.103395
Jones, I., Bisson, M. J., Gilmore, C., & Inglis, M. (2019). Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *British Educational Research Journal*, *45*, 662–680. https://doi.org/10.1002/BERJ.3519
Jones, I., & Davies, B. (2024). 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., & Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics*, *89*, 337–355. https://doi.org/10.1007/s10649-015-9607-1
Kelly, K. T., Richardson, M., & Isaacs, T. (2022). Critiquing the rationales for using comparative judgement: A call for clarity. *Assessment in Education: Principles, Policy & Practice*, *29*(6), 674–688.
Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1–2), 81–93. https://doi.org/10.1093/biomet/30.1-2.81
Kinnear, G., Jones, I., & Davies, B. (2025). Comparative judgement as a research tool: A meta-analysis of application and reliability. *Behavior Research Methods*, *57*(8), 222.
Knight, C., Conn, C., Crick, T., & Brooks, S. (2025). Divergences in the framing of inclusive education across the UK: A four nations critical policy analysis. *Educational Review*, *77*(2), 495–511. https://doi.org/10.1080/00131911.2023.2222235
Korn, G. A., & Korn, T. M. (2000). *Mathematical handbook for scientists and engineers: Definitions, theorems, and formulas for reference and review*. Courier Corporation.
Lambert, B. (2018). *A student's guide to Bayesian statistics*. SAGE Publications Ltd.
Laming, D. (1984). The relativity of 'absolute' judgements. *British Journal of Mathematical and Statistical Psychology*, *37*(2), 152–183.
Lazo, A. V., & Rathie, P. (1978). On the entropy of continuous probability distributions (Corresp.). *IEEE Transactions on Information Theory*, *24*(1), 120–122. https://doi.org/10.1109/TIT.1978.1055832
Leech, T., Gill, T., Hughes, S., & Benton, T. (2022). The accuracy and validity of the simplified pairs method of comparative judgement in highly structured papers. In *Frontiers in Education* (Vol. *7*, 803040). Frontiers Media SA.
Lewis, D. D. (1995). A sequential algorithm for training text classifiers: Corrigendum and additional data. *ACM SIGIR Forum*, *29*(2), 13–19. https://doi.org/10.1145/219587.219592
Lindsay, B. G. (1995). *Mixture models: Theory, geometry, and applications*. Ims.
Magowan, L. (2023). Centre assessment grades in 2020: A natural experiment for investigating bias in teacher judgements. *Journal of Computational Social Science*, *6*(2), 609–653.
Marshall, N., Shaw, K., Hunter, J., & Jones, I. (2020). Assessment by comparative judgement: An application to secondary statistics and English in New Zealand. *New Zealand Journal of Educational Studies*, *55*, 49–71. https://doi.org/10.1007/s40841-020-00163-3

McGaughey, F., Watermeyer, R., Shankar, K., Suri, V., Knight, C., Crick, T., Hardman, J., Phelan, D., & Chung, R. (2022). 'This can't be the new norm': Academics' perspectives on the COVID-19 crisis for the Australian University Sector. *Higher Education Research & Development*, *44*(8), 2231–2246. https://doi.org/10.1080/07294360.2021.1973384

Mentzer, N., Lee, W., & Bartholomew, S. (2021). Examining the validity of adaptive comparative judgment for peer evaluation in a design thinking course. *Frontiers in Education*, *6*, 772832. https://doi.org/10.3389/feduc.2021.772832

Morris, R., Gorard, S., See, B., & Siddiqui, N. (2023). Can a code-based approach to marking and feedback reduce teachers' workload? An evaluation of the flash marking intervention. *Oxford Review of Education*, *50*, 552–569. https://doi.org/10.1080/03054985.2023.2258779

Mussweiler, T., & Epstude, K. (2009). Relatively fast! Efficiency advantages of comparative thinking. *Journal of Experimental Psychology: General*, *138*(1), 1–21.

National Education Union. (2024). Workload advice. https://neu.org.uk/advice/your-rights-work/contracts-and-working-hours/workload-and-working-time/workload-advice

Neyman, J. (1992). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. In *Breakthroughs in Statistics*Springer (pp. 123–150). Springer New York.

Nisbet, I., & Shaw, S. D. (2020). *Is assessment fair?* SAGE Publications Ltd.

Norton, L., Floyd, S., & Norton, B. (2019). Lecturers' views of assessment design, marking and feedback in higher education: A case for professionalisation? *Assessment & Evaluation in Higher Education*, *44*, 1209–1221. https://doi.org/10.1080/02602938.2019.1592110

Norton, S., & Hack, K. (2024). *Framework for enhancing assessment in higher education*. Advance HE.

Office for Students. (2022). *The regulatory framework for higher education in England*. https://www.officeforstudents.org.uk/publications/the-regulatory-framework-for-higher-education-in-england/

Palisse, J., King, D., & MacLean, M. (2023). Does comparative judgement reduce students' perceived cognitive load when evaluating mathematics solutions? In *Proceedings of the Australian Conference on Science and Mathematics Education* (pp. 77–82). https://openjournals.library.sydney.edu.au/IISME/article/view/17348

Pitt, E., & Winstone, N. (2018). The impact of anonymous marking on students' perceptions of fairness, feedback and relationships with lecturers. *Assessment & Evaluation in Higher Education*, *43*(7), 1183–1193. https://doi.org/10.1080/02602938.2018.1437594

Pollitt, A. (2012). Comparative judgment for assessment. *International Journal of Technology and Design Education*, *22*(2), 157–170. https://doi.org/10.1007/s10798-011-9189-x

Quality Assurance Agency for Higher Education. (2018). Uk quality code for higher education: Advice and guidance on assessment. https://www.qaa.ac.uk/the-quality-code/advice-and-guidance/assessment

Raaper, R. (2016). Academic perceptions of higher education assessment processes in neoliberal academia. *Critical Studies in Education*, *57*(2), 175–190. https://doi.org/10.1080/17508487.2015.1019901

Ragolane, M., Patel, S., & Salikram, P. (2024). Ai versus human graders: Assessing the role of large language models in higher education. *Asian Journal of Education and Social Studies*, *50*, 244–263. https://doi.org/10.9734/ajess/2024/v50i101616

Rasooli, A., Zandi, H., & DeLuca, C. (2018). Re-conceptualizing classroom assessment fairness: A systematic meta-ethnography of assessment literature and beyond. *Studies in Educational Evaluation*, *56*, 164–181. https://doi.org/10.1016/J.STUEDUC.2017.12.008

Read, B., Francis, B., & Robson, J. (2005). Gender, 'bias', assessment and feedback: Analyzing the written assessment of undergraduate history essays. *Assessment & Evaluation in Higher Education*, *30*(3), 241–260. https://doi.org/10.1080/02602930500063827

Scharaschkin, A., & Baird, J. (2000). The effects of consistency of performance on a-level examiners' judgements of standards. *British Educational Research Journal*, *26*, 343–357. https://doi.org/10.1080/713651557

Senanayake, C., & Asanka, D. (2024). Rubric based automated short answer scoring using large language models (llms). In *2024 International Research Conference on Smart Computing and Systems Engineering (SCSE)* (Vol. *7*, pp. 1–6). https://doi.org/10.1109/SCSE61872.2024.10550624

Siegel, A., Zarb, M., Alshaigy, B., Blanchard, J., Crick, T., Glassey, R., Hott, J. R., Latulipe, C., Riedesel, C., Senapathi, M., Simon, & Williams, D. (2021). Teaching through a global pandemic: Educational landscapes before, during and after COVID-19. In *Proceedings of the 2021 Working Group Reports on Innovation and Technology in Computer Science Education (ITiCSE-WGR'21)*. https://dl.acm.org/doi/10.1145/3502870.3506565

Sivia, D., & Skilling, J. (2006). *Data analysis: A Bayesian tutorial* (2nd ed.). Oxford University Press.

Spencer, J. R., & Horn, D. (2023). The three most important words in faculty workload: Transparency, transparency, transparency. *The Department Chair*, *34*(1), 1–3. https://doi.org/10.1002/dch.30522

Steedle, J., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*, *29*, 211–223. https://doi.org/10.1080/08957347.2016.1171769

Strathern, M. (2000). The tyranny of transparency. *British Educational Research Journal*, *26*(3), 309–321. https://doi.org/10.1080/713651562

Stuulen, J., Bouwer, R., & van den Bergh, H. (2024). Effects of a comparative feedback method on peer feedback characteristics and revision quality. *L1-Educational Studies in Language and Literature*, *24*, 1–21. https://doi.org/10.21248/l1esll.2024.24.1.671

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286. https://doi.org/10.1037/h0070288

Tierney, R. (2014). Fairness as a multifaceted quality in classroom assessment. *Studies in Educational Evaluation*, *43*, 55–69. https://doi.org/10.1016/J.STUEDUC.2013.12.003

Velasco-Martinez, L. C., & Tojar-Hurtado, J. C. (2019). Transparency in evaluation through the use of rubrics in university subjects. *New Trends and Issues Proceedings on Humanities and Social Sciences*, *6*(1), 54–164. https://doi.org/10.18844/prosoc.v6i1.4166

Wainer, J. (2023). A Bayesian Bradley-Terry model to compare multiple ML algorithms on multiple data sets. *Journal of Machine Learning Research*, *24*(341), 1–34.

Walker, S. (2025). *Trends in assessment in higher education: considerations for policy and practice*. Jisc.

Watermeyer, R., Bolden, R., Knight, C., & Crick, T. (2025). Academic anomie: Implications of the 'great resignation' for leadership in post-COVID higher education. *Higher Education*, *89*, 1215–1233. https://doi.org/10.1007/s10734-024-01268-0

Watermeyer, R., Crick, T., & Knight, C. (2022). Digital disruption in the time of COVID-19: Learning technologists' accounts of institutional barriers to online learning, teaching and assessment in UK universities. *International Journal for Academic Development*, *27*(2), 148–162. https://doi.org/10.1080/1360144X.2021.1990064

Watermeyer, R., Crick, T., Knight, C., & Goodall, J. (2021). COVID-19 and digital disruption in UK universities: Afflictions and affordances of emergency online migration. *Higher Education*, *81*, 623–641. https://doi.org/10.1007/s10734-020-00561-y

Watermeyer, R., Shankar, K., Crick, T., Knight, C., McGaughey, F., Hardman, J., Suri, V. R., Chung, R., & Phelan, D. (2021). 'Pandemia': A reckoning of UK universities' corporate response to COVID-19 and its academic fallout. *British Journal of Sociology of Education*, *42*(5–6), 651–666. https://doi.org/10.1080/01425692.2021.1937058

Willey, K., & Gardner, A. (2010). Improving the standard and consistency of multi-tutor grading in large classes. In *Assessment: Sustainability, Diversity and Innovation. A Conference on Assessment in Higher Education*. https://opus.lib.uts.edu.au/handle/10453/16662

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Gray, A., Rahat, A., Lindsay, S., Pearson, J., & Crick, T. (2026). Rendering transparency to ranking in educational assessment via Bayesian comparative judgement. *Review of Education*, *14*, e70149. https://doi.org/10.1002/rev3.70149