Bayesian Active Learning for Comparative Judgement: A New Paradigm for Educational Assessment

Andy Gray
[redacted]

Submitted to Swansea University in partial fulfilment
of the requirements for the Degree of Doctor of Philosophy

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

30th September 2025

Declarations

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed … [redacted] (candidate)

Date ……30/09/25……

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed [redacted] (candidate)

Date ……30/09/25……

I hereby give consent for my thesis, if accepted, to be available for electronic sharing.

Signed … [redacted] … (candidate)

Date ……30/09/25……

The University's ethical procedures have been followed and, where appropriate, that ethical approval has been granted.

Signed … [redacted] … (candidate)

Date ……30/09/25……

I would like to dedicate this work to my daughter, my partner, and my mother. Thank you for all of your support.

Abstract

Assessment is a cornerstone of education, yet traditional marking methods can be inconsistent, biased, and cognitively demanding for educators. Comparative Judgement (CJ) offers an alternative by ranking student work through pairwise comparisons but faces challenges around transparency, efficiency, and biases in pair selection. This thesis proposes Bayesian Comparative Judgement (BCJ) as a structured, data-driven alternative to address these limitations.

BCJ integrates entropy-driven active learning to select the most informative comparisons, systematically improving ranking accuracy while avoiding the model deterioration seen in traditional CJ. By generating complete predictive rank distributions, BCJ also enables probabilistic grading aligned with assessors' decisions. Experiments using synthetic and real-world datasets, including GCSE essays, demonstrate BCJ's strong performance and efficiency against existing CJ methods. BCJ also introduces methods to quantify assessor agreement, reinforcing reliability and accountability

To enhance performance insights, the thesis extends BCJ to a multi-criteria framework (MBCJ), aligning with rubric-based assessment by evaluating individual learning outcomes (LOs) alongside overall performance. This approach preserves CJ's efficiency while providing granular, criterion-specific insights.

The practical implementation of BCJ and MBCJ is evaluated through studies with professional markers in higher education, focusing on transparency, usability, and trust. Findings show that while traditional marking is familiar, BCJ and MBCJ reduce subjectivity, improve alignment with target rankings, and maintain educational integrity, offering a viable alternative for assessment.

Overall, this research demonstrates how Bayesian methods can refine CJ to enhance transparency, fairness, and efficiency in assessment. The thesis supports adopting structured CJ methods to improve feedback and reduce educator workload, contributing to fairer and more transparent assessment practices across diverse educational contexts.

Acknowledgements

First and foremost, I would like to thank my partner, my daughter, and my mother for their unwavering support and patience throughout this journey.

Secondly, I am deeply grateful to my supervisory team, Dr Alma Rahat, Professor Tom Crick, Professor Jen Pearson, and Dr Stephen Lindsay, for their invaluable guidance, encouragement, and expertise.

Finally, I would like to extend a special thanks to Darren Wallace from CDSM. Without their collaboration with Swansea University, I would not have had the opportunity to undertake this PhD.

Contents

List of Tables

List of Figures

Glossary of Notations

$(\gamma_1, \ldots, \lambda_D)$	Weight vector for D LOs
α, β	Shape parameters of a Beta distribution
$\Gamma(\cdot)$	Gamma function
γ_i	Latent preference (ability) parameter for item i in the BTM
γ_j	Latent preference (ability) parameter for item j in the BTM
$\gamma_{i,d}$	Score of item i for LO d
λ_d	Weight for learning outcome d
$\mathbb{E}[r_i]$	Expected rank of item i
$\mathbf{r}$	Rank vector (initialised to mean rank)
erf$(\cdot)$	Error function (used in Thurstone model)
μ_i	Latent preference score (logit parameter) for item i in the BTM
μ_i	Mean score for item i (Normal model)
μ_j	Latent preference score (logit parameter) for item j in the BTM
$\pi(p_{[i,j]})$	Posterior Beta distribution over performance for pair (i, j)
$\psi(\cdot)$	Digamma function
σ_i	Standard deviation for item i (Normal model)
τ	Kendall's rank correlation coefficient

B Budget (total number of comparisons), $B = N \times K$

$B(\alpha, \beta)$ Beta function

$F(\cdot)$ Cumulative distribution function (CDF)

G List of selected pairs

$H(\cdot)$ Entropy function

I Set of items under comparison

K Budget multiplier for comparisons

$L(\cdot)$ Likelihood function in BTM

m Mean-difference term $m = (\mu_i - \mu_j)/\sqrt{\sigma_i^2 + \sigma_j^2}$

N Number of items being compared

$P(i \succ j)$ Probability that item i is preferred to j

r_i Rank of item i

W List of winners

w Number of wins

x_k Outcome of the kth Bernoulli trial (0 or 1), i.e. a pair-wise comparison

z_a Combinatorial term $\binom{N-1}{a-1}$ for rank probability

${}_2F_1(\cdot)$ Gaussian hypergeometric function

List of Acronyms

A-Level Advanced Level

ACJ Adaptive Comparative Judgement

AI Artificial Intelligence

AfL Assessment for Learning

AoL Assessment of Learning

BCJ Bayesian Comparative Judgement

BTM Bradley–Terry Model

CJ Comparative Judgement

CSE Certificate of Secondary Education

EAP Expected Agreement Percentage

GCSE General Certificate of Secondary Education

HCI Human–Computer Interaction

JSD Jensen–Shannon Divergence

LO Learning Outcome

ML Machine Learning

MAP Mode Agreement Percentage

MBCJ Multi-Criteria Bayesian Comparative Judgement

MC Monte Carlo

MCMC Monte Carlo Markov Chain

MCP Mixture of Component Preferences

MCR Mixture of Component Ranks

O-levels The General Certificate of Education (GCE) Ordinary Level

PM Principal Moderator

QMC Quasi–Monte Carlo

SAT Standard Assessment Test

SSR Scale Separation Reliability

UI User Interface

UX User Experience

Contributing Publications

	Contributing Chapter	Publication
1	Chapter 2	Using Elo rating as a metric for Comparative Judgement in educational assessment: Proceedings of the 6^{th} ACM International Conference on Education and Multimedia Technology
2	Chapter 2 & 3	A Bayesian active learning approach to Comparative Judgement within education assessment: Computers and Education: Artificial Intelligence
3	Chapter 2 & 4	Bayesian Active Learning for Multi-Criteria Comparative Judgement in Educational Assessment [Journal paper is in preparation]
4	Chapter 2 & 5	Rendering Transparency to Ranking in Educational Assessment via Bayesian Comparative Judgement: BERA Review of Education on Transparency in Assessment

Table 1: List of publications and their associated chapters developed by the author during the preparation of this thesis.

Ethical Considerations

For the real data used in Chapter 3, we only received anonymised pairwise comparison data (i.e. no identifying information or exact grades, or scores, for the items, the original text, or information about the judges, were provided) from the authors. Hence, we were merely the users of anonymised data, and since it is owned by the authors, we do not have the privilege to share it. Interested readers should reach out to them for this data. Ethics approval for the use of the secondary data was approved by the Faculty of Science and Engineering ethics committee at Swansea University (Research Ethics Approval Number: 1 2024 9700 8643).

The experiment carried out in Chapter 5 was approved by the Faculty of Science and Engineering ethics committee at Swansea University (Research Ethics Approval Number: 1 2023 7465 6926).

Chapter 1

Introduction

Subjectivity, bias and inequity influence the overall judgement on a pupil's performance [1], leading to fundamental questions such as: "is assessment fair?" [2]. Inconsistency in teachers' predictions of student grades is widespread in UK schools and colleges. In 2019, only 21% of students obtained the grades predicted by their teachers [3], while in 2011, 42–44% of teacher-estimated grades over-predicted by at least one grade, and 7–11% under-predicted [4].

The COVID-19 pandemic forced reliance on predicted grades across educational settings and contexts. Its impact was immediate and profound [5, 6, 7, 8, 9, 10], and its long-term consequences are still not fully manifested [11, 12, 13, 14]. We will likely continue to experience a "new normal" for education [15, 16, 17, 18, 19, 20], particularly in the domain of educational assessment [21, 9, 22, 23, 24]. During the pandemic, student grades were based on teachers' assessments in England and Wales (two of the four nations of the UK, with separate education systems), resulting in record-high grades for GCSE and A-level students. However, with the announcement of the 2022 A-level results, $80{,}000$ fewer students received A and A^* grades compared to 2021, a fall from 19.1% achieving A^* in 2021 to 13.5% in 2022—ultimately bringing grades back in line with pre-pandemic results [25].

There is an extensive corpus of work that focuses on using intelligent and/or data-driven approaches in a variety of educational settings and contexts [26, 27, 28, 29, 30]. In particular, for predicting student performance and retention, we have seen broad application of data mining and learning analytics [31, 32], as well as machine learning (ML), collaborative filtering, recommender systems, and artificial neural networks [33, 34, 35].

However, these approaches raise increasingly complex and interconnected social, ethical, legal, and digital/data rights issues [36, 37, 38], especially when viewed through a pre- and post-pandemic lens [39]. In an emerging educational policy context, the potential for disempowering educators and undermining their expertise in supporting learning and progression via formative and summative assessment approaches is also problematic.

1.1 Motivations & Objectives

Within education in the UK, since the introduction of national curricula and Key Stages (KS) 1, 2, and 3 in 1988, assessments have been used to rank students' attainment [40]. The introduction of the national curriculum and KS 1, 2, and 3 brought a stronger emphasis on assessment, allowing teachers to provide structured feedback to support student improvement [41, 42]. As a result, new assessment approaches emerged, notably Assessment of Learning (AoL) and Assessment for Learning (AfL) [43, 42, 44]. However, traditional marking, also known as absolute marking, remains time-consuming and laborious, contributing to teacher workload and stress, particularly when schools enforce strict marking policies with specific time constraints. Additionally, biases can arise, as teachers may unconsciously base grades on prior student performance rather than the specific assessment itself.

This thesis aims to develop novel artificial intelligence (AI) and ML techniques to enhance Comparative Judgement (CJ) in e-learning and e-assessment. By leveraging data from digital learning platforms, the goal is to support educators in assessing qualitative work more effectively. Traditional automated assessment tools struggle with qualitative responses due to their subjective nature, often leading to inconsistencies and inefficiencies in grading. This research seeks to bridge that gap by developing an ML framework that can identify meaningful patterns in qualitative assessments while ensuring fairness and reliability.

A key aspect of the project is the integration of CJ methods to improve grading consistency and transparency. Educators will be actively involved in shaping the system, helping to refine the model's ability to recognise key assessment attributes. Through an interactive learning process, the system will use educator feedback to enhance its understanding of high-quality submissions, allowing it to highlight areas that require closer attention. By adopting active learning strategies, the model will prioritise uncertain cases, making

the assessment process more efficient without constraining educators to rigid, automated grading structures.

Additionally, this research prioritises trust and usability, ensuring that both educators and students feel confident in the system. By involving stakeholders throughout the design process, the project will address concerns around bias and transparency. The final tool will not only support educators in marking and feedback but will also help ensure alignment across different assessors, improving assessment reproducibility. Ultimately, this work aims to create a system that enhances learning outcomes, supports constructive alignment, and informs future developments in education technology.

1.2 Contributions

The main contributions of this work can be seen as follows:

- We develop a comprehensive Bayesian framework for pairwise preference-based assessment that derives an analytical expression for the full predictive rank distribution of any item; updates pairwise preference densities and rank distributions via Bayesian inference as new comparison data arrive; introduces a novel active learning strategy using predictive entropy to select the most informative next pair; proposes a probabilistic grading approach based on predictive rank distributions; defines two new reliability metrics—Mode Agreement Percentage (MAP) and Expected Agreement Percentage (EAP)—under Beta priors to quantify assessor agreement and identify controversial pairs; and demonstrates through extensive synthetic experiments that our BCJ active learning framework with entropy-based selection achieves statistically superior or equivalent performance across all tested configurations.

- We extend the BCJ framework to a multi-criteria setting (MBCJ), introducing methods to approximate overall ranks and predictive uncertainties from pairwise comparisons for each learning outcome (LO), and to derive LO-specific predictive rank distributions; we define a holistic entropy measure to guide active selection of the next pair for evaluation; and, for the first time, demonstrate through experiments on real assessment data that MBCJ performs as well as BCJ while providing finer granularity in item preferences and LO-specific rankings.

- We demonstrate how BCJ enhances the transparency of CJ by addressing criticisms of traditional CJ through a structured, explainable process that tracks decision-making and quantifies ranking uncertainty; conduct a comprehensive evaluation comparing traditional BCJ, and MBCJ against standard marking in real-world assessment contexts, providing practical insights into their relative performance; and analyze educators' perspectives—via quantitative analyses and expert discussions—on fairness, workload, and usefulness, ensuring the research is grounded in authentic teaching and assessment practices.

Taken together, this research presents a significant advancement in assessment methodology by introducing the BCJ and MBCJ frameworks, which improve the reliability, transparency, and interpretability of grading in education.

1.3 Chapter Outlines

In Chapter 2, we present the literature and explore the limitations of traditional assessment methods, such as absolute marking and rubric-based grading, which often suffer from inconsistencies, biases, and high cognitive demands on educators. CJ is introduced as a potential alternative, leveraging pairwise comparisons to improve reliability and reduce marking burden. While CJ offers advantages, challenges such as transparency, time requirements, and potential biases in adaptive approaches are discussed. The section also examines alternative ranking methods, including Bayesian approaches.

In Chapter 3, we introduce our novel approach to CJ, Bayesian CJ. The proposed method enhances the process of educational assessment through CJ using Bayesian active learning. We focus on utilising an ML model that improves assessment accuracy by learning from human judgements. In particular, the method aims to reduce the number of comparisons needed to evaluate educational content, thus making the process more efficient. The approach utilises Bayesian methods to refine the model iteratively, improving the precision of assessments and reducing the inherent subjectivity in traditional evaluation methods. This research offers a framework for integrating ML techniques into educational assessment, potentially improving both the fairness and efficiency of grading. Additionally, we address the limitation of quantifying assessor agreement by deriving agreement levels, thereby enhancing transparency in the assessment process.

In Chapter 4 we extend our novel BCJ approach to handle multi-criteria decision marking. The proposed innovative approach aims to bridge the gap between holistic assessment methods and the need for detailed, criterion-based feedback in education. Traditional CJ allows assessors to evaluate work holistically through pairwise comparisons, leading to reliable rankings but lacking specific criterion-based insights. Building upon the Bayesian CJ (BCJ) framework, this study extends it to handle multiple independent learning outcome components as defined by rubrics. This extension enables both overall and component-wise predictive rankings, complete with uncertainty estimates. We introduce an entropy-based active learning method to select the most informative comparisons for assessors, enhancing the efficiency of the assessment process. Experiments on synthetic and real datasets demonstrate the effectiveness of this approach.

In Chapter 5, we explore the application of BCJ to enhance transparency in educational assessment. Traditional CJ methods, while effective for ranking student work, often lack transparency and detailed feedback. BCJ introduces a probabilistic approach that incorporates prior information, enabling clearer decision-making and quantification of uncertainty. The study further extends this to MBCJ, which evaluates multiple learning outcomes independently, offering a more granular and transparent ranking system. Through an experiment with professional markers in higher education, this chapter demonstrates that BCJ and MBCJ improve consistency compared to absolute marking while maintaining fairness. The findings suggest that BCJ and MBCJ could serve as reliable alternatives to traditional assessment methods, particularly in contexts demanding high levels of transparency.

Having established the motivation, objectives, and challenges surrounding traditional assessment methods, the next chapter delves into the theoretical and empirical foundations that underpin this research. Chapter 2 provides a comprehensive literature review, exploring the evolution of assessment practices, the emergence of CJ, and the computational models that support ranking methodologies. This foundation is essential for understanding the innovations proposed in subsequent chapters.

Chapter 2

Literature Review & Background

Our target in this thesis is well-rooted within the educational space, aiming to provide educators with an alternative method of conducting CJ, enabling transparency and accurate ranking of students' work while addressing current CJ issues and keeping the marking burden on teachers as low as possible. Therefore, we will explore multiple areas spanning different scientific research spaces. These include computer science for looking into the multiple different methods that can be used to implement alternative versions to use ML, and human-computer interaction (HCI) to be able to ensure that what we create is beneficial to the main target of these solutions (educators) as well as the social science field to ensure that what we are aiming to produce is rooted within the requirements of the education field and expected manner that the educational domain expects.

In this chapter we explore the literature in four interconnected domains: section 2.1 focuses on teaching and learning with a emphasis on traditional assessment methods and their limitations, section 2.2 CJ and its theoretical underpinnings, 2.2.2 algorithmic approaches for ranking in CJ, and section 2.2.3 ML innovations to address known limitations.

2.1 Teaching and Learning

While there can be multiple reasons why educators assess students, assessments serve a purpose for both the teacher and the student in the process. These include: giving feedback to teachers and learners; providing motivation and encouragement; boosting the pupils' self-esteem; a basis for communication; a method to evaluate a lesson, training

method, scheme of work, or curriculum; and to entertain the students [43]. Additionally, the assessment process creates opportunities to rank students, ultimately allowing schools to select and filter students, allocate them to a particular pathway or educational direction, or discriminate between students for a given reason [43].

There are four main categories of assessment in the UK. These are diagnostic, formative, summative, and national assessments [42, 43]. However, it is essential to note that national assessments are not used in everyday aspects of teaching and learning. The term' national assessment' is used to represent critical examinations, such as SATS (Standard Assessment Tests), GCSE (General Certificate of Secondary Education), and A-level (Advanced Level) examinations, taken at the end of the qualifications [43].

Teaching is becoming increasingly target-oriented and evidence-based in the UK. For example, the Office for Standards in Education (Ofsted) in England carry out inspections and conducts teacher performance reviews [45, 46]. Therefore, there is an emphasis on rigour, and consequently, teachers are encouraged to provide written summative feedback [47, 48]. While verbal feedback has been proven to be just as effective for students' learning as written feedback, it is challenging to record and provide evidence to governing bodies [49, 50, 51]. Therefore, a teacher will mark students' work based on a rubric of key criteria that match certain levels to produce a grade and then provide personalised feedback to a student, explaining what they have done well and what they need to improve on. However, this approach to marking is very time-consuming and generates a substantial marking burden for the teachers [52, 53]. Additionally, When marking is conducted over multiple sessions, assessors may become inconsistent in their marking and feedback. Such temporal drift can unintentionally introduce systematic bias into assessment outcomes, as markers' internal standards shift over time [54], considering how the student has performed all year round rather than in that moment of the assessment [55].

2.1.1 Assessment

Education is widely regarded as a foundational mechanism for personal and societal development. As the saying goes, "give a man a fish and you will feed him for a day, but teach him to fish and he will not be hungry anymore." However, it was not until 1918 that education, as most people in England and Wales know it, began to take its current form [56].

In the intervening decades, education policy in England and Wales widened participation and reshaped teaching practice. The Education Act 1944 established free secondary education and introduced a tripartite system of grammar, technical, and secondary modern schools [57]. Post-war reforms encouraged broader curricula and more child-centred methods. At the same time, qualifications such as O-levels (The General Certificate of Education (GCE) Ordinary Level) and CSEs (Certificate of Secondary Education) began to standardise assessment at the secondary level [57]. These developments laid the groundwork for the comprehensive national curriculum and assessment structures that followed in 1988 [40].

For much of the twentieth century, teaching remained largely teacher-centred, focused on transmitting knowledge to students. It was not until 1988, under the Education Reform Act, that assessment became a formal and integral part of schooling through the introduction of the national curriculum in England and Wales [40].

As the curriculum was rolled out, statutory assessments were introduced to education between 1991 and 1995. Key Stage 1 first, followed by Key Stages 2 and 3, respectively [41, 42]. Only for the core subjects of English, Mathematics, and Science had the assessments been first introduced. The first assessments in Key Stage 1 were a range of cross-curricular tasks to be delivered in the classroom, known as SATs. However, the complexity of using these meant that more formal assessments quickly replaced them [41, 42]. The assessments in Key Stages 2 and 3 were developed using more traditional tests.

To allow teachers to assess students' attainment, taking tests became the primary assessment method in Key Stage 3. While assessments were the primary means, educators were also able to evaluate their students using other methods against the targets set for attainment within the national curriculum [42]. The teacher and assessment outcomes were used on a scale with key learning milestones expected at different ages. A key stage level indicated the result for the students' progress. The model was used throughout the next few years until 2005, when the role of tests in KS1 was downgraded to just being an internal support tool for teachers. Then, in 2008, the government decided to remove tests in KS3 [42].

This model continued, with minor adjustments to reflect the changing content of the National Curriculum, up to 2004. From 2005, the role of the tests got downplayed at Key Stage 1, with tests being used only internally to support teacher assessment judgements

[58]. Further changes occurred in 2008 when the government announced that testing in Key Stage 3 would be scrapped altogether [59].

However, with a change of government, the Conservative Party taking power from the Labour Party brought about new changes to how education is focused and pedagogy methods would be conducted. In 2014, the system of attainment levels was removed, creating the educational shift of "Assessing without level" [60]. However, within schools, it was being referred to as 'life after levels'. This was the follow-up to the changes in the national curriculum in 2013 [60]. The changes within the national curriculum brought a greater focus on more traditional-style GCSE academic subjects, while reducing the emphasis on perceived technical labour-style jobs. The new curriculum's direction has created a greater emphasis on final exam outcomes at the GCSE and A-level stages.

Traditional marking, an action where a marker assigns a score and/or grade to a piece of work, is the default approach to assessment in education [61, 62]. This practice is also known as absolute marking, and this is the term used throughout the remainder of this paper. In this method, teachers assign marks based on fixed criteria or rubrics, aiming to gauge a student's performance objectively. Despite its widespread use, this approach has been critiqued for issues related to consistency, bias, transparency, and, crucially, the cognitive demands it places on educators [63].

On inconsistency, even when a grading rubric is in place, researchers have shown that different teachers can interpret the same criteria in varying ways, leading to discrepancies in scoring [64, 65]. Additionally, biases, such as those based on a student's previous performance, personality, or even handwriting, can unintentionally affect marks. Scharaschkin *et al.* [65] highlights the "halo effect", where a teacher's perception of a student's past achievements influences their grading of current work. This effect can lead to an overestimation or underestimation of a student's true capabilities, which is particularly problematic in high-stakes assessments.

Biases can also stem from factors like teacher fatigue, stress, and subjective preferences [66]. Research has shown that teachers' grading decisions can be influenced by non-academic student characteristics, such as gender, socioeconomic background, or ethnicity, leading to systematic patterns of disadvantage even when attainment is held constant [67]. For example, different tutors within a large marking team may apply the same marking criteria in subtly different ways, leading to systematic variation in the marks awarded [68].

These biases undermine the fairness of assessments, which can affect students' opportunities and their trust in the educational system [69].

It should be noted that while anonymisation is often deemed as a definitive way to combat such biases and improve trust in the system, evidence suggests that it may not necessarily be effective (see, for example, [70]). Also, it is not always feasible to use anonymisation for all types of assessments, for instance, for presentation.

Teachers in England work an average of 54 hours per week, while school leaders work more than 60, according to the UK Government's Department for Education (DfE)'s workload survey [71] (*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) [72]. Recognising this, the DfE has emphasised the need to reduce school workload, providing a toolkit of practical resources to support reductions [73].

In 2019, the average self-reported working hours for all teachers and middle leaders was 49.5 hours per week, a reduction of 4.9 hours from 2016. Primary teachers and middle leaders reported working an average of 50 hours per week in 2019, down from 55.5 hours in 2016, while secondary teachers and middle leaders reported a decrease from 53.5 hours to 49.1 [71]. However, primary teachers continue to work longer hours than their secondary counterparts, though the gap has narrowed from 2 hours per week in 2016 to 0.9 hours 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 [74]. 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 [74].

This marking workload leads to stress, poor wellbeing [75], and has also been shown to be associated with systematic shifts in grading behaviour under repetitive marking conditions [76]. As student numbers increase, absolute marking becomes increasingly difficult to sustain at scale, motivating interest in automated approaches [77]. 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 judgments [78], which can contribute to variability in grading outcomes across educational contexts [79].

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 [11, 13, 14]. 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. [80, 81, 82].

Furthermore, 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 particularly concerning in high-stakes situations, where grades can have a significant impact on future educational and career opportunities. Transparency is vital to ensuring trust in the assessment process, yet absolute marking methods often fall short in providing the necessary clarity [83].

The HE sector is also compounded by structural sector-wide issues in part arising from the COVID-19 pandemic [11, 17, 84]. 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 [81, 9, 82]. 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 [80]. Human essay marking, even with rubrics, often produces substantial inter-marker variability, suggesting that reliability and fairness of assessment can be deeply compromised under current practice standards [85].

2.1.2 The Purpose of Assessment, Marking and Feedback in Education

As we have established, assessments became a staple of the UK educational system in 1988. While the term assessments is not usually defined, the word 'assess' is typically associated with measuring, determining, evaluating, and judging [43].

While there can be multiple reasons why educators assess students, assessments serve a purpose for both the teacher and the student in the process. These include giving feedback to teachers and learners, providing motivation and encouragement, boosting pupils' self-esteem, serving as a basis for communication, a method to evaluate lessons, a training method, and a scheme of work on the curriculum [43]. Additionally, the assessment also creates other opportunities to rank students; a method to select and filter students, allocate students a particular pathway or educational direction, or as a way to discriminate or choose between students for a given set of reasons [43].

2.1.3 Traditional Methods of Assessment and Feedback

There are four main categories of assessment. These are diagnostic, formative, summative, and national assessments [43, 42]. However, it is essential to note that national assessments are not used in everyday aspects of teaching and learning. This term refers to the critical exams, such as SATs, GCSEs, and A-level exams, taken nationally. Therefore, we will focus on the other three main ones.

Diagnostic assessment is also referred to as pre-testing [43]. Educators use this technique to acquire a base level of knowledge of the students they have inherited. This method is good for showing the progress of attainment over time by having an initial base test. Teachers can then show how well the students have progressed over time with their improvements over the term. This base assessment also provides the teacher with crucial information, the current ability of every student's knowledge. Through knowing this current level of knowledge, teachers can adapt the coming lessons and provide suitable differentiation and scaffolding within the lessons to allow each student to succeed as much as possible. However, during the author's time as an educator, it was also observed that the technique was employed to create baseline narratives. Teachers used them to show that the student's knowledge was not at the expected level when inherited by the teacher at meetings or performance management reviews. Therefore, being used as a counter-act measure tool by the teacher, if they find themselves being accused of letting

the students' performance slip, they try to counter-act by implying the students were not at the required level in the first place.

The second method, formative assessment, is also known as 'assessment for learning (AfL)' [43, 42]. This method has become one of the main tools for a teacher in terms of assessment and feedback. AfL allows educators to assess students' understanding of a topic in real-time during a lesson, without relying on a summative assessment. As a result, the technique allows the teacher to spend more or less time on the topic, depending on whether the students understand it or not, even if they had planned to spend more or less time on the topic. Therefore, ensuring that the teaching is not happening for the sake of teaching. Thus, the emphasis is less on measurements and more on actual learning. AfL can involve using several techniques, such as teacher assessment through in-class questions and marking books, as well as self-assessment and peer assessment, where students evaluate each other's work [43].

AfL has many values for teachers and students. Within Black and William's paper, 'Inside the black box' [44], it was discovered that AfL provides massive learning gains, especially with the low attainer groups. Black and William found that AfL and the use of peer assessment raised motivation and self-esteem across the board, but particularly among low attainers. The addition of peer assessment is extra valuable to the students. This form of feedback is effective because it will most likely be presented to the students in a manner with which they are more familiar, in terms of language and wording. Therefore, in a way that makes more sense to them and has the most impact on their learning [86, 44].

The two key ways that teachers can gain insights from AfL are in questioning and marking. Questioning, also referred to as formative questioning, aims to assess what students in the classroom know about the current topic being discussed or taught, thereby improving learning [43]. However, for this to be effective, students will need an appropriate 'wait time' [87]. A 'wait time' is the term used to ensure that the student, when asked a question, has sufficient time to formulate their thoughts and answer, as the aim is not to catch them out but to gather what they currently understand. Formative questioning is also effective when allowing students to discuss amongst themselves and then answer the teacher. Therefore, allowing them to consolidate with peers to check if they understand the topic before delivering it to the teacher. A student is more likely to say they do not know than give a wrong answer and look silly in front of their peers, a technique known as think–pair–share. Other effective methods, which do not require students to discuss

between themselves, include no-hands up, where the teacher selects who answers rather than relying on volunteers; show-me boards, where every student writes an answer on a small whiteboard and reveals it simultaneously so the teacher can gauge understanding across the class; and traffic-light systems, in which learners indicate their confidence or level of understanding by showing red, amber, or green cards (or similar signals) [88].

Formative marking is the term used when teachers assess students' work and provide some form of feedback, whether it be 'two stars and a wish' or more standard approaches of giving direct feedback. The overall aim is to enable the teacher to assess the student's current level of knowledge, identify areas for improvement, and provide feedback on both their strengths and areas for development. Giving feedback on areas for improvement is essential, whether the student is at a C/4 or an A*/9. The constant feedback, regardless of the student's level, is an educator's aim to ensure that their students can improve. However, it is crucial that the feedback is taken on board and actioned for formative marking to be effective. Otherwise, it is more of a summative action [44, 89]. To combat this, educators would usually allow students time within a lesson, after the feedback has been given, to review their work and make changes in a different colour. This process is referred to as DIRT (Directed/dedicated. Improvement/Independent. Reflection.) [90].

The third method is a summative assessment, also known as 'assessment of learning' (AoL) [43]. This type of assessment happens at the end of a teaching unit or topic. It is used to gain insights into what the students have learnt within the subject covered or the course. Its purpose is to give a student a mark, grade or ranking. Usually, this is the grade that is mainly focused on, as it is the metric that will have the most impact on the school in terms of league performance tables regarding GCSE and A-level results. This assessment method is used to acquire a snapshot of the students and allows the teacher to perform 'what if' scenarios, such as if they were to take the test now, what would they get? Educators can determine if students need to attend intervention, are performing as expected, or are even performing better by reviewing the results. With so much riding on these results, particularly for schools and teachers' performance management reviews, considerable emphasis is placed on predicting the final results for students. We have seen it put much pressure on the teachers and the students, and ultimately creates a very stressful environment, which is not the best environment for learning.

2.1.4 The Negative Aspects of Marking and Feedback Methods

While marking and feedback are essential in the classroom, they also have some negative aspects. Currently, debates are happening about who formative assessment is really for [43]. Are these assessments for the students designed to allow them to improve knowledge, or are they more for the schools to predict their work and where the students will be, come exam time? Are they there to demonstrate to external bodies, such as Ofsted, that the school is being rigorous? Or are they for teachers to justify possible results based on results for their performance management reviews?

Additionally, as teachers might have had a KS4 (GCSE) class for two to three years when assessing and conducting summative assessments, they may not see the student's work entirely at face value. The teacher's personal bias might influence their assessment based on how the student has performed over the year or even years. For example, if one student has been consistently friendly, well-behaved, and has completed the required work, the teacher might provide a higher grade for that student. However, a teacher may be inclined to assign a lower grade to a student who has exhibited disruptive behaviour throughout the year. Consequently, the quality of that student's work may not be fully recognised, and the assessment may be less accurate, as extraneous factors influence the judgement.

Because schools often have several teachers delivering the same subject, a process of moderation is required to ensure that marking and grading are applied consistently. Moderation seeks to verify that, for example, a Distinction* (star) awarded by teachers A, B or C reflects the same agreed standard of quality. However, this process can present challenges. Teachers may interpret the mark scheme differently, focusing on varying aspects of students' work. Moderation and standardisation are intended to identify and resolve such inconsistencies, but organisational dynamics can complicate matters.

Consider a scenario in which five teachers share responsibility for the same year group and qualification. One is the designated subject lead, another is a classroom teacher; others include an assistant principal, a vice principal, and the head of faculty. Within the context of the qualification, the subject lead holds the highest authority; however, in the broader school hierarchy, they rank below several colleagues. This dual structure can invite 'office politics': senior staff outside the qualification remit may assert their own interpretation of the mark scheme, even when it diverges from the lead. The subject lead must then decide whether to challenge a more senior colleague or concede to preserve

collegial relations—often the latter, given the complexities of school management. Such dynamics risk undermining consistency in marking and the reliability of awarded grades.

Another drawback to absolute marking is the requirement for personalised feedback for students. To allow them to develop, students must have personalised areas where they need to improve. However, in controlled assessments, teachers can give feedback, but it can not be personalised. It has to be generic, but most schools' policies require the feedback to be personalised, creating a conflict between the exam board's requirements and the school's requirements based on Ofsted's expectations. The situation makes a moral and ethical decision. They are likely to be reprimanded by the school if they do not provide the feedback, but can be investigated for malpractice by the exam board if they catch them giving the feedback or providing too much guidance to the student.

Such situations highlight the difficult balance teachers must maintain between supporting pupils and complying with examination regulations.

Where a teacher's actions are judged to constitute malpractice—whether by providing excessive guidance, failing to follow assessment procedures, or otherwise breaching the Joint Council for Qualifications (JCQ) regulations—the consequences can be significant for both staff and students.

For pupils, penalties range from the loss of marks for a single component to disqualification from an entire qualification or, in serious cases, a ban on sitting future examinations with the awarding body [91, 92].

Teachers and invigilators may receive formal warnings, be removed from assessment responsibilities, or be referred to professional regulators such as the Teaching Regulation Agency, which can impose restrictions or prohibition orders [91]. Schools themselves may be subjected to enhanced monitoring or even the withdrawal of exam-centre status if systemic malpractice is found [91].

These measures aim to protect the integrity of national assessments but place educators in a position where well-intentioned feedback can risk serious professional and institutional consequences.

Another factor is when a summative assessment has occurred within a learning sequence, students are usually presented with a grade and feedback. This feedback and mark could be for the end of unit exams or homework, for example. While the teachers want students to focus on the feedback given to help them improve, students focus on the results and will naturally rank order themselves. The UK government has attempted to

try and resolve this by removing levels in KS3. However, when KS4 focuses on the final summative assessment, their actual GCSE exams, a provided grade is hard not to offer. Therefore, it is vital to make sure that feedback is acted upon once given.

Finally, a big issue in regards to marking and providing feedback is time. It takes a long time to score a students' work and then give feedback to the students. It is also a very tedious task that a teacher might not do in one sitting. Therefore, with many potential variables in play, the marking of the points award per each exam question, for example, might not be the same. There is also a massive burden that is placed upon the teacher while trying to mark.

Consequently, it is challenging to ensure that consistency and fairness play a part in the marking. However, the enormous burden placed upon the teacher can be very draining. It can then affect the quality of the teacher's delivery within the lesson, especially with the stress aspects that are placed upon them regarding how quickly the feedback needs to be returned to the students.

These issues—bias, inconsistency, and the marking burden of absolute marking highlight the need for more structured tools that can improve grading consistency. One widely adopted approach is rubric-based marking, which aims to formalise assessment criteria and enhance objectivity.

2.1.5 Rubric Marking

Rubrics, grounded in explicit criteria and specific expectations, provide a systematic framework for evaluating students' knowledge and their ability to apply it effectively [93]. Recent research in Education for Sustainable Development shows a growing shift toward holistic educational frameworks that integrate cognitive, emotional, and social learning dimensions, highlighting broader interest in holistic approaches within certain educational research communities [94]. Rubric marking has become increasingly used by educators [95]. Sambell *et al.* [96] emphasise that assessment should be designed to promote meaningful student learning, supported by clear criteria and timely, constructive feedback. Although they do not refer to rubrics explicitly, their principles, particularly those stressing transparency of expectations and shared understanding of standards, align closely with the pedagogical purposes of rubric-based assessment.

A marking rubric, also referred to as a scoring rubric, is a tool that delineates the expectations for an assignment by listing criteria and describing levels of quality. It offers a

	70–100%	60–69%	50–59%	40–49%	20–39%	0–19%
Implementation **(50%)**	Excellent to outstanding implementation of web development techniques, far beyond class scope. Detailed commenting, flawless structure, strong conventions.	Good to very good implementation. Sound use of semantic markup, attention to detail, good structure and commenting.	Fair implementation. Core techniques used, with some scope for refinement. Comments and structure could be improved.	Basic implementation with modest errors. Minimal commenting and poor presentation.	Poor implementation with significant errors. Limited understanding and poor structure.	Very limited implementation, demonstrating little to no understanding.
Additional Requirement **(25%)**	Excellent to outstanding response. Wide range of techniques and non-trivial problem solving. Iterative process evident.	Good to very good response. Sound understanding and thoughtful application of techniques.	Fair response with appropriate techniques used. Some errors present, but minor.	Basic response with limited depth. May contain major to modest errors.	Poor response. Limited scope and significant errors.	Very limited response showing little to no understanding.
Documentation **(25%)**	Excellent to outstanding documentation. Highly precise, well-structured, and concise. Strong reflection and self-evaluation.	Good to very good documentation. Clear, concise, and technically accurate. Minor improvements possible.	Fair documentation. Acceptable structure. All sections present but lacking depth. Technical terminology could improve.	Basic documentation with key information but lacking depth. One or more limited sections.	Poor documentation. Badly structured and fragmented. Missing sections.	Very limited documentation with missing content and no structure.

Figure 2.1: An example marking rubric for a level 4 undergraduate module offered at the Bath Spa University, UK. It provides an overview of the quality required to achieve a certain grade (along the columns), based on different criteria (along the rows) for the assessment as designed by the assignment owner. Here, the criteria are Implementation, Additional requirements, and Documentation.

clear and objective method to assess student work, including essays, group projects, creative endeavours, and oral presentations. Rubrics can be employed for any assignment in a course, or for any way in which students are asked to demonstrate what they've learned.

Rubric marking has solidified its role as a structured evaluative method within educational assessment, offering a systematic approach to gauging student performance. The deployment of scoring rubrics is backed by extensive research, which underscores their reliability, validity, and impact on learning outcomes. The consistency of rubric-based assessments is well-supported, particularly when they are analytic, subject-specific, and bolstered by exemplars and rater training, as noted by Jonsson *et al.* [97]. Although rubrics are not inherently valid, their validity can be enhanced through a comprehensive validity framework during the rubric validation process [97]. The explicit criteria provided by rubrics facilitate feedback and self-assessment, promoting learning and improving instruction [97]. When clear and focused, descriptive rubrics yield high-quality information [98], which contributes to the overall positive impact of rubrics on student performance.

While the effects on self-regulation of learning are mixed, there is evidence supporting a positive correlation between the use of rubrics and motivation to learn [98].

There are two primary types of rubrics: holistic and analytical [99]. Holistic rubrics consolidate all scoring criteria into a single scale [100]. Holistic rubrics consolidate multiple aspects of performance into a single overall judgement, which can make them relatively efficient to apply but also introduces challenges when student work demonstrates uneven performance across different criteria [100, 101]. Analytic or descriptive rubrics, by contrast, decompose performance into multiple, explicitly defined criteria, providing more detailed diagnostic information but increasing the structural complexity of rubric design and use [100, 97]. Evidence suggests that scoring consistency is strongest when analytic rubrics are subject-specific and supported by exemplars and rater training, although such rubrics require careful construction and validation [97]. Analytic scoring schemes include explicit descriptions of expected learning outcomes and detailed criteria defining performance levels for each dimension of assessment [102].

Rubrics have several advantages, such as providing clarity and consistency in grading [97]. They offer clear expectations and grading criteria to students, which can help them understand what is required to excel in an assignment [103]. While rubrics do not inherently reduce marking time, they can streamline decision-making by clarifying criteria and supporting consistent judgement [97]. Furthermore, rubrics can provide students with informative feedback on their strengths and weaknesses, allowing them to reflect on their performance and work on areas that need improvement [104]. Rubrics also encourage learners to develop critical thinking about their own scores and work [104]. However, rubrics also have their drawbacks. The language of rubrics is not always as straightforward as it is supposed to be, which adds to their complexity [105]. The lower scale may use negative terms to describe student performance, which may discourage the learners. Some opponents of rubrics believe they are more subjective than letter grades [105].

In higher education, rubrics have been recognised for enhancing student self-assessment, self-regulation, and understanding of assessment criteria [103]. However, some students perceive rubrics as restrictive and associate them with increased stress related to assessments [103]. The involvement of students in the design and implementation of rubrics is essential for their success [103]. In primary education, particularly in the teaching and assessment of mathematical reasoning, rubrics have been found to improve teachers' diagnostic skills and indirectly influence their use of formative feedback [106]. However,

the direct effects on student self-assessment are more apparent than the effects on student outcomes, highlighting the need for further research into the mediated effects of self-regulation and self-efficacy [106].

Educators have pinpointed both effective and ineffective practices in rubric use [107]. Good practices include the standardisation of evaluation methods and transparency, while ineffective practices involve vague descriptions and a lack of clear marking ranges [107]. These insights are instrumental in developing rubrics that ensure fair and consistent marking. Although less common, rubrics have also found their place in program evaluation, aiding in the transformation of data, characterisation of organisational functioning, and derivation of evaluative conclusions [108]. In teacher education, rubrics are emphasised for their ability to communicate expectations, scaffold learning, and support the development of critical thinking skills [109]. It is essential for teacher candidates to understand how to create, administer, and use rubrics effectively [109].

Empirical research on the use of analytic rubrics in higher education, highlighting their role in supporting reliable and consistent assessment [110]. While the reliability of rubrics is supported by evidence, the impact on student learning necessitates further robust evaluation [110]. Ultimately, rubrics are invaluable tools in educational assessment, with their effectiveness contingent upon their design, implementation, and the context in which they are used. The potential of rubrics is vast, yet challenges remain that require ongoing research to understand and address fully. When effectively implemented, rubric marking can significantly enhance the reliability and validity of assessments, positively influencing student learning and performance. However, the actual impact of rubric marking varies depending on specific contexts and implementations, and it is influenced by factors such as the clarity of criteria, assessor training, and the feedback provided to students. These general pros and cons underscore the need for a nuanced application of rubrics in educational settings.

While rubrics offer improved structure and clarity, they are not immune to subjectivity, nor do they fully resolve concerns around feedback quality, student motivation, or fairness. To address these broader concerns, particularly those related to openness and trust in assessment, the concept of transparency requires further examination.

2.1.6 Transparency in Assessment

Traditional marking, also known as absolute assessment, is the dominant form of grading in education. In this method, teachers assign marks based on fixed criteria or rubrics, aiming to gauge a student's performance in an absolute sense. Despite its widespread use, this approach has been critiqued for issues related to consistency, bias, transparency, and the cognitive demands it places on educators.

One major concern with absolute marking is its potential for inconsistency. Even when a grading rubric is in place, studies have shown that different teachers can interpret the same criteria in varying ways, leading to discrepancies in scoring [64, 65]. Additionally, biases—such as those based on a student's previous performance, personality, or even handwriting—can unintentionally affect marks. Scharaschkin *et al.* [65] highlights the "halo effect," where a teacher's perception of a student's past achievements influences their grading of current work. This effect can lead to an overestimation or underestimation of a student's true capabilities, which is particularly problematic in high-stakes assessments.

Biases can also stem from factors like teacher fatigue, stress, and subjective preferences. Research indicates that teachers' grading decisions are often influenced by non-academic factors, even if they are subconscious. For example, teachers may give higher marks to work that aligns more closely with their own views or personal standards [68]. These biases undermine the fairness of assessments, which can affect students' opportunities and their trust in the educational system [69].

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 [111, 112] – perhaps more so following the COVID-19 pandemic [5, 113], and the inexorable rise and widespread impact of AI in education [29, 114, 115]. 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 [116, 74]. Fairness in classroom assessment is a multifaceted construct, encompassing consistency, transparency, and the procedural legitimacy of marking decisions, all of which are central to students' acceptance of assessment outcomes [117, 118]. A lack of transparency in marking undermines stakeholder confidence in the fairness and legitimacy of assessment outcomes [119].

Furthermore, when grading is inconsistent or biased, students may feel frustrated or sceptical about the fairness of the assessment process [120, 69]. This lack of transparency is particularly concerning in high-stakes situations, where grades can have a significant impact on future educational and career opportunities. Transparency is vital to ensuring trust in the assessment process, yet absolute marking methods often fall short in providing the necessary clarity [83].

Absolute marking is time-consuming and demanding for teachers. Teachers' mental resources can be depleted by the effort required to assess large volumes of work accurately and consistently [75]. 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 [121].

The high burden of marking often leads to fatigue, which can reduce the accuracy and consistency of grades over time. Teachers under pressure to meet grading deadlines may resort to shortcuts, such as relying more heavily on first impressions or previously formed opinions about students' abilities [76]. These shortcuts, while understandable, further compromise the reliability and validity of assessments. Additionally, time constraints and workload pressures can lead to "marking fatigue", where teachers' grading quality deteriorates as they progress through a stack of assessments [76].

In terms of time, absolute marking can place significant demands on teachers, often requiring them to spend hours outside of class reviewing and grading work [116]. 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 [75]. The inefficiency of absolute grading systems is a significant drawback, especially when compared to alternative assessment methods that can streamline the process [122].

Traditional absolute marking in education presents several challenges, including inconsistencies, bias [123], a lack of transparency [124], and the negative impact on teachers' wellbeing [75]. These issues highlight the need for more reliable, transparent, and efficient assessment methods that support both educators and students in the learning process.

In UK HE, assessment transparency has become a core component of institutional accountability, quality assurance, and external scrutiny, with the Office for Students, the Quality Assurance Agency, and Advance HE all emphasising the need for defensible

and equitable assessment practices [125, 126, 127]. These policies and practices have increasingly emphasised inclusive, fair, and flexible assessment strategies, with growing attention to the challenges posed by digital transformation, evolving pedagogies, and changing student needs [128]. Thus, despite sector-wide initiatives to enhance and improve the 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 [129]. Bloxham (2009) [130] argues that rigorous moderation procedures in UK higher education create a substantial workload burden for markers, which constrains assessment choices and slows feedback. She suggests that this burden may limit the feasibility of moderation and calls for alternative approaches that reduce marker workload while preserving assessment rigour. However, new and innovative methods must be understood in relation to these governance frameworks that shape their adoption and perceived legitimacy.

While rubrics offer structure, they still suffer from subjectivity and inconsistency. This has led to interest in alternative assessment frameworks, such as CJ, which we explore in the next section.

Together, the challenges of consistency, bias, workload, and lack of transparency within absolute marking approaches have driven interest in alternative assessment frameworks. One such method is CJ, which moves away from assigning absolute scores in favour of relative comparisons and offers the potential to alleviate some of the burdens placed on educators.

2.2 Comparative Judgement

Prospect theory suggests that humans are typically more effective at making *relative* rather than *absolute* judgements [131]. In educational assessment, this observation underpins the method of CJ, which has been proposed as an alternative to conventional rubric-based marking [132, 133]. In CJ, assessors (e.g. teachers) are repeatedly presented with pairs of scripts and asked to decide which is of higher quality. A ranked order of items is then inferred from the set of pairwise outcomes using a statistical model of CJ, commonly the Bradley–Terry model (BTM) [134], following Thurstone's foundational work on the *law of comparative judgement* (LCJ) [135]. In CJ, an overall rank order can be estimated from relative pairwise comparisons alone, rather than from absolute scores. Educational

research has further suggested that such comparative decisions may reduce assessors' cognitive demands, as judging between two pieces of work avoids the need to assign absolute marks across multiple criteria [136].

Thurstone's LCJ formalised the idea that judgements are made on a latent psychological continuum rather than in absolute measurement terms [137]. Early applications in psychometrics and psychophysics included comparing perceived weight intensities, attitude extremity, or relative size [138, 139]. Pollitt and colleagues introduced and popularised CJ within educational contexts, where raters judge scripts holistically against the construct of interest [133, 140, 132, 141].

CJ presents assessors with two items—in this case, two apples—and asks a single holistic question (e.g. "Which apple has the darker colour?"). Rather than assigning grades, the assessor makes a relative choice between the pair. This is, in essence, the process of CJ (as illustrated in Figure 2.2). However, instead of this being a one-time process, it is repeated across many different pairs and, where appropriate, multiple assessors. This simple two-at-a-time decision is cognitively lighter than absolute marking and exploits humans' strength in making relative judgements, while still enabling robust, model-based inferences about overall quality or intensity (here, perceived darkness) [142, 132].

Let v_a and v_b denote the latent quality parameters (location on a common scale) of items A and B. In the BTM (consistent with Thurstone's Case V under a logistic link), the log-odds that A beats B is

$$\log \text{odds}\big(A \text{ beats } B \,\big|\, v_a, v_b\big) \;=\; v_a - v_b\,. \tag{2.1}$$

Equivalently, the probability that A is judged better than B is

$$P\big(A \text{ beats } B \,\big|\, v_a, v_b\big) \;=\; \frac{\exp(v_a - v_b)}{1 + \exp(v_a - v_b)}\,. \tag{2.2}$$

Here, the left-hand side represents the observable outcome (the probability that A is preferred), while v_a and v_b are latent values to be estimated from the full set of pairwise outcomes. The difference $v_a - v_b$ thus provides a natural, transitive basis for ranking items [131].

In practice, CJ proceeds by repeatedly selecting pairs, eliciting the winner, and updating the estimated rank after each comparison until a pre-specified budget of judgements is exhausted. A generic procedure is summarised in Algorithm 1, and a visual overview is shown in Figure 2.3.

Figure 2.2: An example of a pairwise comparison: two items (e.g. apples) presented side by side for judgement, with the aim for the judge to make a judgement. In this example, which apple has the darker colour?

Algorithm 1 Standard comparative judgement procedure.

Inputs.

N: number of items
K: multiplier controlling the budget of pairwise judgements
I: set of items

Steps.

$B \leftarrow N \times K$ ▷ Compute the budget (number of pairs).
$G \leftarrow \langle\,\rangle$ ▷ Initialise the list of selected pairs.
$W \leftarrow \langle\,\rangle$ ▷ Initialise the list of observed winners.
$\mathbf{r} \leftarrow \left(\frac{N}{2}, \ldots, \frac{N}{2}\right)^\top$ ▷ Initialise ranks with the mean rank.
for $b = 1 \rightarrow B$ **do**
    $(i, j) \leftarrow \text{SelectPair}(I)$ ▷ Select a pair of items.
    $G \leftarrow G \oplus \langle (i, j) \rangle$ ▷ Append the selected pair.
    $w \leftarrow \text{DetermineWinner}(i, j)$ ▷ Record the chosen winner.
    $W \leftarrow W \oplus \langle w \rangle$ ▷ Append the winner.
    $\mathbf{r} \leftarrow \text{GenerateRank}(G, W)$ ▷ Update rank estimates.
end for
return r

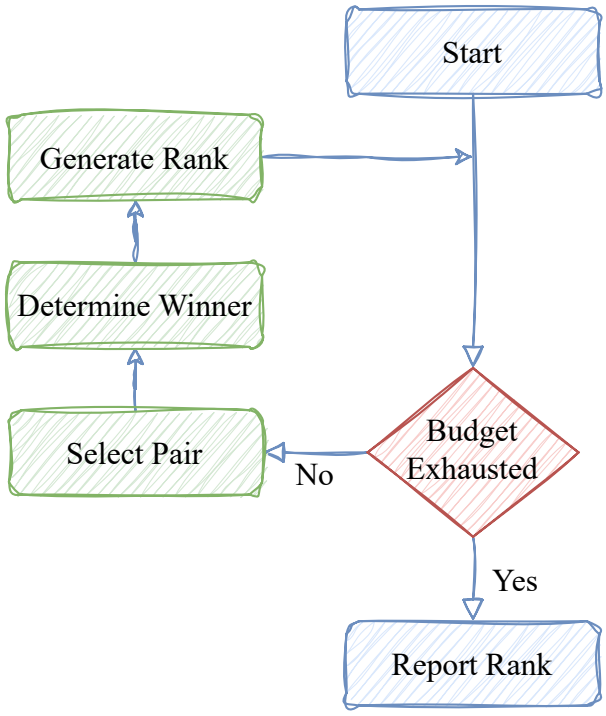


Figure 2.3: A flow chart of the CJ process. Starting from the set of items to be ranked, pairs are selected up to a pre-specified budget and shown to assessors, who choose a winner on each trial. A statistical model (e.g. BTM) updates the rank estimates after each decision is made. Once the budget is exhausted, the final rank order is reported. Green boxes highlight core CJ components that vary across methodological implementations.

In conventional rubric-based marking, a teacher assigns absolute scores to each script independently, ideally without being influenced by prior knowledge about the students. In practice, perfect anonymisation and memory suppression are difficult to achieve, and prior expectations or classroom experiences can unintentionally influence outcomes [143, 144]. By basing decisions on pairwise comparisons against a holistic construct, comparative judgement avoids the need for absolute score assignment and analytic decomposition. Prior research suggests that this approach can improve scoring consistency and may reduce some sources of bias inherent in absolute marking [132, 141]. From a cognitive load

perspective, such pairwise judgements reduce element interactivity relative to absolute marking, which may lower intrinsic cognitive demands on assessors [145].

CJ is simple in design yet statistically robust: deviations from model predictions can be detected and evaluated, and estimated item locations are placed on a common latent scale, enabling meaningful comparison across pairs [131, 146]. However, practical challenges remain. Ofqual has raised concerns that the quality of the inferred rank order can deteriorate if the number of judgements is insufficient, and that the method can appear less transparent to stakeholders unfamiliar with its statistical underpinnings [147]. Moreover, while adaptive comparative judgement (ACJ) is designed to reduce the number of required comparisons without sacrificing accuracy, the adaptive selection mechanism may introduce its own biases [148]. Consequently, there is ongoing interest in "pure" CJ, as well as in methods that enhance efficiency without compromising validity.

A common measure of CJ reliability is the *Scale Separation Reliability* (SSR) [148, 149, 131]. Conceptually, SSR reflects the ratio of true-score variance to the variance of estimated scores derived from the observations; see Verhavert *et al.* for a detailed treatment [150]. In practice, SSR depends on the chosen CJ model (e.g. BTM) and its uncertainty estimates, and can be difficult to compute in settings where the true-score variance is unknown or must be approximated. Despite these challenges, multiple studies report high reliability for CJ-derived rank orders. For example, across 16 exercises (1998–2015), correlations between CJ outcomes and rubric-based grades ranged from 0.73 to 0.99 [151]; values of 0.70 and above are typically interpreted as indicative of strong agreement [152].

A key practical issue is determining when to stop collecting judgements. No universally accepted, model-agnostic convergence criterion provides a natural stopping rule. As a result, CJ exercises are often run on a fixed budget—e.g. a minimum of 10 judgements per script [153]. This creates a tension between reliability, transparency, and efficiency: too few judgements risk instability in the inferred rank; too many increase cost and time. Developing principled stopping rules and allocation strategies that minimise the number of interactions while maintaining accuracy and transparency remains an important open research problem.

CJ leverages humans' strength in making relative judgements to produce robust, model-based rank orders from pairwise comparisons [135, 134, 131]. It offers potential gains in reliability, fairness, and cognitive efficiency over absolute marking [132, 141], yet faces practical challenges related to transparency, budgeting of judgements, stopping criteria,

and the risk of bias introduced by adaptive mechanisms [147, 148]. Addressing these challenges motivates methods that retain the conceptual simplicity of CJ while reducing the number of required interactions and clarifying uncertainty for end users.

While CJ addresses several shortcomings of absolute marking, its perceived opacity and the limited communication of statistical uncertainty have constrained its broader acceptance [147]. Empirical evidence supports the efficiency of relative judgements [142] and the practical consistency of CJ outcomes [151], yet methodological clarity remains uneven. We therefore develop *Bayesian Comparative Judgement (BCJ)* to (i) render the inferential process explicit, (ii) provide principled uncertainty quantification for ranks and pairwise win probabilities, and (iii) inform more efficient pair-selection policies. The next section formalises rank generation from paired comparisons (Algorithm 1, line 10); we then address pair selection (line 6).

2.2.1 Pairing Selection Methods

The literature presents that the main way to generate a ranking preference is to generate combinations of items to present to a user. There are three main approaches to presenting the judge with an item, which are explained in this section. The user then selects which one they think is better. This result is then tracked, and another pair is presented until the desired number of comparisons has been reached. The results are then calculated using the BTM, producing a final rank. However, Ofqual has stated that if the number of possible pairs becomes too large compared to the optimum number, then the final ranking becomes less effective; however, determining this optimal number of ranks is unknown [147]. Therefore, we propose a new approach that employs Bayesian ML to calculate rankings, providing the marker with greater transparency and insight into how the model generates its rankings.

The number of samples is determined by the number of items N and the K factor, with 10 being the recommended minimum, which would result in 100 comparisons if the sample size is 10 [154]. We will also use different K factors to determine the impact on overall performance, using values of 5, 10, 20, and 30.

2.2.1.1 Random

The random approach uses a method where every pair presented to the user is done at random. There is no intrinsic method behind it other than each time a pair is required, as

the number of pairs already presented is lower than the $N \times K$ target has been reached, and a new pair at random is displayed. This can cause real-world issues, as the same pair may be presented to the user, but this is very unlikely, especially as N increases in size. This is used as a worst-case scenario within our experiment comparisons. This method has been proposed within the literature [154, 132].

2.2.1.2 No Repeating Pairs

This is another approach used in current methods that incorporates an element of randomness, but no repeating pairs occur [154], unless all possible pairs have been seen and the budget hasn't been reached. Then the process will start again. This ensures that all n items are seen the same number of times, but what item is compared against what item is decided randomly. This prevents identical pairs from being presented to a user. Still, it does have an element of stochasticity, as each time a CJ cycle is performed, the items that will be compared against are, to a degree, decided randomly. However, by ensuring that all items are seen the same number of times, it provides a balance between all items in terms of their ability to gain insights into their ranks.

2.2.1.3 Adaptive Comparative Judgement

ACJ is a variant of CJ in which assessors compare pairs of student work to establish a quality ranking, rather than using predefined rubric criteria [131, 155]. Instead of presenting random pairs throughout, ACJ employs a statistical algorithm—typically a form of the Bradley–Terry model—to prioritise comparisons between submissions of similar estimated quality, updating rankings as judgements accumulate [156, 157]. This adaptive scheduling aims to reduce the number of comparisons needed while maintaining or improving reliability, expressed through SSR, and is promoted as scalable for large or complex assessments [131].

However, these reliability claims warrant scrutiny. Simulation studies have shown that ACJ's adaptive algorithm can inflate SSR, sometimes producing high reliability estimates even with random data [148, 156]. This raises concerns that the reported robustness may reflect algorithmic effects rather than genuine assessor consistency [156]. In addition, the holistic nature of ACJ may not align with contexts that require criterion-referenced evidence or fine-grained skill assessment, and some validity arguments rely on circular comparisons with rubric-based marking [158, 159, 160].

Further research has highlighted practical drawbacks. Because the adaptive algorithm focuses comparisons on pieces of work with similar provisional rankings, certain scripts are judged repeatedly, which can increase the number of judgements required in those regions and introduce potential sources of bias in the resulting scale [132, 148]. Consequently, some in the CJ community now favour random pairing, judging it at least as effective and less prone to these artefacts [153, 154].

2.2.2 Bradley-Terry Model

In this thesis, we adopt the Bradley–Terry model (BTM) to derive rankings from pairwise comparisons of student essays, a standard approach in CJ [132, 161, 162, 131, 163]. The model posits that each item has a latent quality score, and the probability of one item being preferred over another depends on the relative magnitudes of these scores.

Let $\mu_i \in \mathbb{R}$ denote the latent utility for item i. For two items X and Y, the BTM assumes

$$\text{logit}(P(X \succ Y)) = \mu_X - \mu_Y, \tag{2.3}$$

so the log-odds that X beats Y equals the difference in their latent utilities [164]. Equivalently, with the probability parameterisation $\gamma_i = \exp(\mu_i) > 0$, we obtain

$$P(i \succ j) = \frac{\gamma_i}{\gamma_i + \gamma_j}. \tag{2.4}$$

Unless stated otherwise, ties are ignored (binary outcomes). Variants that explicitly allow ties are available; see [165].

Let $I = \{1, \ldots, N\}$ index the N items, and let $\omega_{[i,j]}$ denote the number of times item i was preferred over item j (with $\omega_{[i,i]} = 0$). Assuming independence across pairings, the log-likelihood of the performance vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)^\top$ is

$$L(\boldsymbol{\gamma}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \Big[\omega_{[i,j]} \ln(\gamma_i) - \omega_{[i,j]} \ln(\gamma_i + \gamma_j) \Big]. \tag{2.5}$$

Because the model is scale-invariant, we identify the parameters by enforcing $\sum_{i=1}^{N} \gamma_i = 1$.

We maximise (2.5) using the minorisation–maximisation (MM) algorithm [134]. Define $\Omega_i = \sum_{j=1}^{N} \omega_{[i,j]}$ (total wins for i) and $n_{ij} = \omega_{[i,j]} + \omega_{[j,i]}$ (total comparisons between i and j). The kth iterate updates as

$$\gamma_i^{k+1} = \frac{\Omega_i}{\sum_{j \neq i} \frac{n_{ij}}{\gamma_i^k + \gamma_j^k}}. \tag{2.6}$$

Followed by normalisation to enforce the simplex constraint:

$$\gamma_i^{k+1} \leftarrow \frac{\gamma_i^{k+1}}{\sum_{j=1}^{N} \gamma_j^{k+1}}. \tag{2.7}$$

Under standard regularity conditions, these iterates converge to the maximiser of (2.5) [134].

After convergence, items are ranked by sorting $\boldsymbol{\gamma}$ in descending order. Using 1-based indexing, we extract the rank as

$$r_{i \in I} = (N + 1) - \text{argsort}(\boldsymbol{\gamma}), \tag{2.8}$$

where argsort($\boldsymbol{\gamma}$) returns item indices ordered by γ_i ascending. For presentation, γ_i may be rescaled (e.g. multiplied by 100) without affecting the induced order. The process in Equation (2.8) can be repeated to generate the complete rank vector in line 10 of Algorithm 1.

2.2.3 Bayesian Approaches

2.2.3.1 Bayesian Inference

In this thesis, we employ *Bayesian inference* as the core inferential framework for modelling and updating beliefs about latent qualities from pairwise judgements. Bayesian modelling provides a principled way to represent uncertainty, combine prior knowledge with observed data, and obtain full posterior distributions over unknown parameters [166, 167]. Bayes' theorem gives the general update rule for parameters $\boldsymbol{\theta}$ given data $\boldsymbol{Y}$:

$$\underbrace{p(\boldsymbol{\theta} \mid \boldsymbol{Y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{Y} \mid \boldsymbol{\theta})}^{\text{likelihood}} \overbrace{p(\boldsymbol{\theta})}^{\text{prior}}}{\underbrace{p(\boldsymbol{Y})}_{\text{marginal likelihood}}}. \tag{2.9}$$

Suppose a jar contains red/blue balls with an unknown red proportion $H \in [0, 1]$. A neutral prior (e.g. uniform) encodes initial uncertainty. Each draw (red/blue) provides

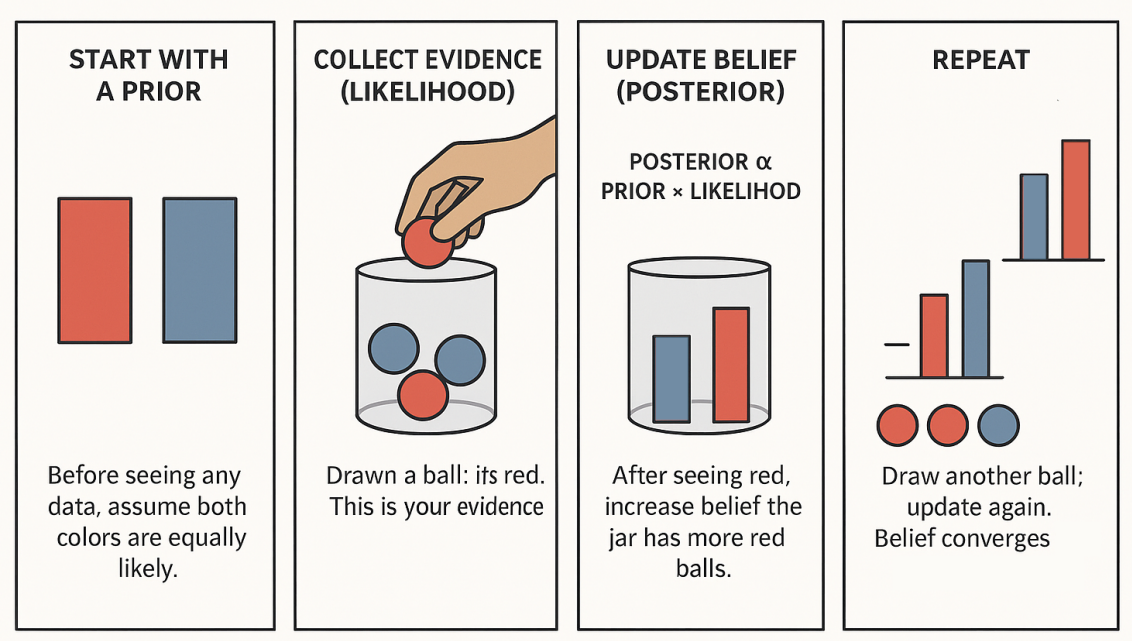


Figure 2.4: An illustration of sequential Bayesian updates: prior → likelihood (new evidence) → posterior.

evidence via the likelihood, and Bayes' theorem updates the belief to the posterior. Repeating the process concentrates the posterior around the true H while retaining a quantified uncertainty.

It is often convenient to write Bayes' theorem up to a proportionality constant,

$$p(X \mid Y, I) \;\propto\; p(Y \mid X, I)\, p(X \mid I), \tag{2.10}$$

where I denotes background information. The omitted denominator $p(Y \mid I)$ does not depend on X and can be recovered by normalisation.

Let H denote the (unknown) probability of drawing a *red* ball. A broad, non-informative prior that reflects minimal assumptions is the uniform density on $[0, 1]$ [168]:

$$p(H \mid I) = \begin{cases} 1, & 0 \le H \le 1, \\ 0, & \text{otherwise}, \end{cases} \tag{2.11}$$

and, assuming independent draws, the likelihood for observing R red balls in N selections is (proportional to) the binomial form [168]:

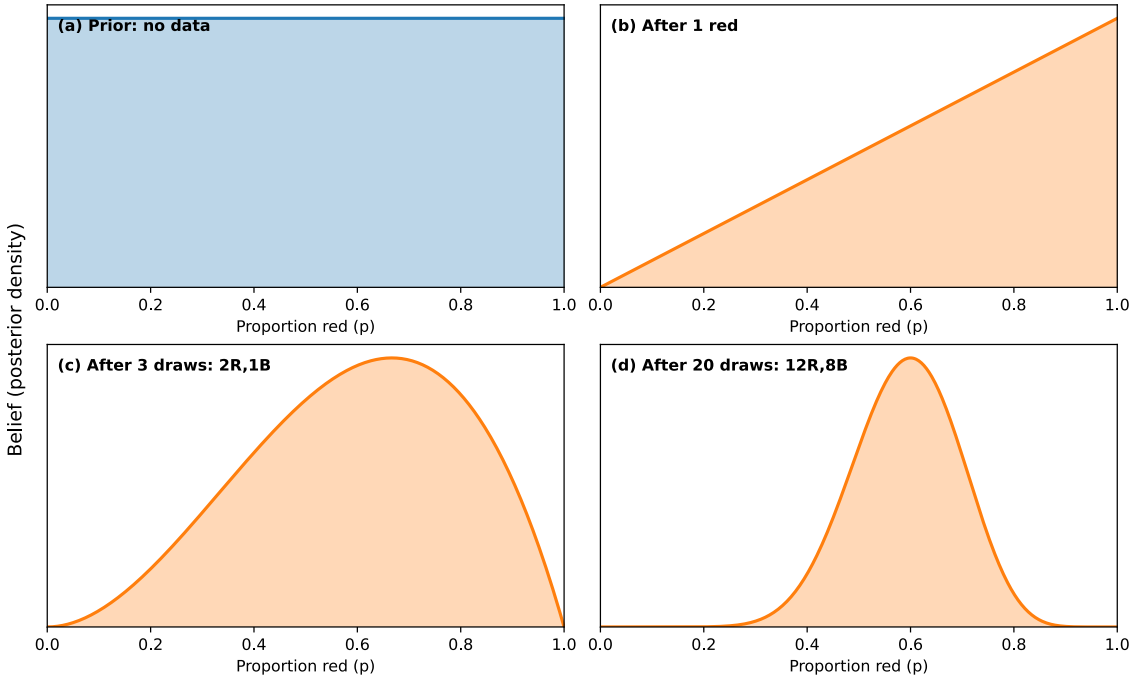


Figure 2.5: A conceptual illustration: prior (uncertain), updates with data (likelihood), and a posterior that concentrates near the true value as evidence accrues.

$$p(\{\text{data}\} \mid H, I) \;\propto\; H^R (1-H)^{N-R}. \tag{2.12}$$

Bayes' rule then yields the posterior (up to normalisation):

$$p(H \mid \{\text{data}\}, I) \;\propto\; p(\{\text{data}\} \mid H, I)\, p(H \mid I), \tag{2.13}$$

with the (rarely needed) normalising constant recoverable from

$$\int p(Y \mid X, I)\, dY = 1, \qquad \int_0^1 p(H \mid \{\text{data}\}, I)\, dH = 1. \tag{2.14}$$

In CJ, each comparison between two items is a *binary* outcome: one item is preferred over the other. Therefore, each observation can be modelled as a Bernoulli trial. Let $Y_{ij} \in \{0, 1\}$ denote that item i is preferred over item j ($Y_{ij} = 1$ if $i \succ j$, else 0). By complementarity, $P(j \succ i) = 1 - P(i \succ j)$. Therefore, we can assume

$$Y_{ij} \mid \pi_{ij} \;\sim\; \text{Bernoulli}(\pi_{ij}), \qquad \text{logit}(\pi_{ij}) = \mu_i - \mu_j,$$

where $\mu_i \in \mathbb{R}$ is the latent quality (utility) of item i. This is the Bayesian counterpart to the Bradley–Terry link. We place weakly informative priors on the latent utilities, e.g. $\mu_i \sim \mathcal{N}(0, \sigma^2)$ with an identifiability constraint (e.g. $\sum_i \mu_i = 0$ or $\mu_1 = 0$). Given a set of pairwise observations $\{Y_{ij}\}$ collected under the usual independence assumptions, the posterior over $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)^\top$ is obtained via Bayes' theorem:

$$p(\boldsymbol{\mu} \mid \{Y_{ij}\}) \;\propto\; \left[\prod_{i<j} \pi_{ij}^{Y_{ij}} (1-\pi_{ij})^{\,1-Y_{ij}}\right] \;\times\; \prod_{i=1}^{N} p(\mu_i),$$

with $\pi_{ij} = \text{logit}^{-1}(\mu_i - \mu_j)$. Inference can be performed using MCMC or variational methods; posterior means/medians of μ_i (or their induced probabilities) yield rankings, while credible intervals provide principled uncertainty quantification around ranks and pairwise win probabilities.

Relative to purely frequentist estimation, Bayesian approaches naturally incorporate prior information (e.g. calibration pieces or historical performance), and they provide full uncertainty on parameters and ranks, and also support online updating as new comparisons arrive, which are all desirable properties in CJ settings.

Bayesian treatments of paired comparisons and CJ are well established. Tsukida *et al.* [169] survey Bayesian analyses for paired comparison data, including Bradley-Terry generalisations. Wainer [170] developed a Bayesian Bradley–Terry model, using MCMC to compare multiple ML algorithms across datasets; their posterior predictive checks indicate that a more complex Davidson tie model is unnecessary for their data within CJ specifically. De Maeyer [171] demonstrates Bayesian CJ using the `pcFactorStan` framework on argumentative writing from the D-PAC project, leveraging Bayesian tools to analyse CJ data rather than drive the live CJ process itself.

Notation remark. Throughout, we write $p(\cdot)$ for probability mass/density functions (context clarifies which), and use $\succ$ to denote "is preferred to" (e.g. $i \succ j$).

2.2.3.2 Entropy Calculations

Entropy is a fundamental concept in information theory, originally introduced by Shannon to quantify the uncertainty in a probabilistic system [172]. It measures the average amount of information produced by a stochastic source of data. Mathematically, for a discrete random variable X with probability mass function $P(x)$, the entropy $H(X)$ is defined as:

$$H(X) = -\Sigma_{x \in \mathcal{X}} P(x) \log P(x) \tag{2.15}$$

Where $\mathcal{X}$ is the set of possible outcomes, and the logarithm is typically taken to base 2, yielding the result in bits. Entropy reaches its maximum when the probability distribution is uniform, indicating maximal uncertainty.

In ML, entropy is commonly used in classification tasks as a measure of impurity or unpredictability in a dataset. For instance, in decision tree algorithms such as ID3 and C4.5, entropy guides the selection of attributes by quantifying the information gain resulting from a split [173]. Information gain is defined as the difference in entropy before and after a split:

$$\text{Information Gain}(S, A) = H(S) - \Sigma_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \tag{2.16}$$

Where S is the set of training examples and A is the attribute being considered. This use of entropy enables the algorithm to prefer features that most effectively reduce uncertainty about class labels.

Beyond decision trees, entropy is also used in regularisation methods, reinforcement learning, and Bayesian modelling, where controlling or maximising entropy can lead to improved exploration or model robustness [174, 175].

In Bayesian inference, entropy is intimately linked to the concept of uncertainty in posterior distributions. The Beta distribution, often used as a conjugate prior for Bernoulli and binomial likelihoods, offers a useful context in which to explore entropy. The Beta distribution with parameters α and β is defined as:

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \tag{2.17}$$

where B(α, β) is the Beta function. The entropy H of a Beta distribution is given by [176]:

$$H(Beta(\alpha, \beta)) = \log B(\alpha, \beta) - (\alpha - 1)\psi(\alpha) - (\beta - 1)\psi(\beta) + (\alpha + \beta - 2)\psi(\alpha + \beta) \tag{2.18}$$

Here, $\psi(\cdot)$ is the digamma function. The entropy increases with increasing uncertainty in the distribution—for instance, when α and β are both close to 1, resulting in a flatter distribution. In ML applications, this entropy can quantify epistemic uncertainty, especially in probabilistic classifiers and Bayesian deep learning models [177].

Understanding entropy and its role in quantifying uncertainty lays the foundation for designing effective active learning strategies, particularly in ranking tasks where uncertainty plays a central role.

2.2.3.3 Active Learning for Ranking from Pairwise Comparisons

Building on the role of entropy as a measure of uncertainty (Section 2.2.3.2), we now consider its use in *active learning* for ranking from pairwise comparisons. The goal is to infer a (partial or total) order over n items using outcomes of pairwise queries of the form "does i beat j?". Acquiring all $\binom{n}{2}$ comparisons is typically infeasible in large-scale or human-in-the-loop settings, motivating adaptive strategies that select the *most informative* comparisons to minimise the number of queries needed for an accurate ranking.

Early work treated pairwise outcomes as deterministic and noise-free. Under geometric structure—specifically, when items admit a low-dimensional Euclidean embedding consistent with an underlying ranking—Jamieson and Nowak [178] showed that adaptivity can dramatically reduce the number of required comparisons compared with naive sorting ($O(n \log n)$) or random sampling (which may approach all $\binom{n}{2}$ pairs). Their algorithms exploit this structure so that query complexity scales with the embedding dimension d (up to logarithmic factors), highlighting the value of adaptively targeting informative pairs [178, 179].

Deterministic assumptions, however, are often unrealistic in human judgement settings, where responses exhibit noise due to cognitive biases, fatigue, and ambiguity. Classic results in behavioural decision research (e.g. Tversky [180]) show that preferences can be intransitive and probabilistic. This motivates *stochastic* comparison models such as Bradley–Terry–Luce (BTL) and Thurstone, where the outcome of comparing i and j is a Bernoulli random variable with success probability determined by latent item scores. In this setting, active learning seeks queries that most reduce uncertainty about the ranking, with analyses of query complexity and algorithms that adaptively select pairs to minimise estimation error under noise [179].

Most active ranking methods under BTL/Thurstone adopt a frequentist view, estimating latent scores via maximum likelihood and using confidence intervals or upper-confidence bounds to guide which comparison to query next. While effective, such procedures can under-represent parameter uncertainty in small-sample regimes. A Bayesian approach,

by contrast, represents uncertainty explicitly via posterior distributions and naturally incorporates prior knowledge, which is advantageous for principled exploration and data efficiency—particularly early in the querying process.

The (differential) entropy of a Beta distribution is:

$$H(\text{Beta}(\alpha,\beta)) \;=\; \log B(\alpha,\beta) \;-\; (\alpha-1)\psi(\alpha) \;-\; (\beta-1)\psi(\beta) \;+\; (\alpha+\beta-2)\psi(\alpha+\beta), \tag{2.19}$$

where $B(\cdot,\cdot)$ is the Beta function and $\psi(\cdot)$ is the digamma function [176]. This entropy serves as a principled acquisition signal: pairs with larger $H(\text{Beta}(\alpha_{ij},\beta_{ij}))$ correspond to greater posterior uncertainty about θ_{ij}, and are thus more informative to query.

Practically, we select the next query via a simple *maximum-entropy* rule

$$(i^\star, j^\star) \;=\; \arg\max_{(i,j)} \; H(\text{Beta}(\alpha_{ij},\beta_{ij}))\,, \tag{2.20}$$

Or, when computational budget allows, by maximising the *expected information gain*

$$(i^\star, j^\star) \;=\; \arg\max_{(i,j)} \; I(Y_{ij};\theta_{ij} \mid \mathcal{D}) \;=\; \underbrace{\mathbb{H}[Y_{ij} \mid \mathcal{D}]}_{\text{predictive entropy}} \;-\; \underbrace{\mathbb{E}_{\theta_{ij}\sim p(\cdot\mid\mathcal{D})}\mathbb{H}[Y_{ij} \mid \theta_{ij}]}_{\text{aleatoric component}}, \tag{2.21}$$

which decomposes uncertainty into epistemic (reducible, via the posterior over θ_{ij}) and aleatoric (irreducible outcome noise) components. Both criteria prioritise comparisons that are expected to reduce posterior uncertainty about the ranking the most.

Consistent with intuition, $H(\text{Beta}(\alpha,\beta))$ is largest when the posterior is diffuse and near-uniform over $(0,1)$ (e.g. $\alpha \approx \beta \approx 1$), and decreases as the posterior concentrates (large $\alpha+\beta$) or becomes highly skewed. Thus, as evidence accumulates for a pair, its entropy drops, and the acquisition rule naturally shifts budget to more uncertain pairs.

Active ranking benefits from modelling and *targeting* uncertainty. Deterministic models illustrate the value of adaptivity under structural assumptions [178], while stochastic formulations better reflect human variability and measurement noise [180, 179]. A Bayesian, entropy-guided strategy provides a principled and data-efficient mechanism to choose comparisons, explicitly quantifying uncertainty and focusing queries where they matter most [176]. Unlike Bayesian optimisation, which aims to locate the maximiser of an unknown function, our objective is to recover a reliable ranking with as few pairwise queries as possible; entropy serves as the unifying signal that directs exploration toward the most informative comparisons.

2.3 Human-Computer Interaction

Human-Computer Interaction (HCI) played a crucial role in our research, as our primary aim was to support teachers with their marking. To ensure our study remained relevant to their needs, we incorporated HCI principles into both the design and execution of our experiments. HCI, often associated with user interface (UI) design, is a field concerned with how humans interact with computers, striving for an optimal balance of complexity and simplicity in these interactions [181]. While initially focused on traditional computing, HCI has since broadened its scope to encompass various aspects of information technology design, making it highly relevant to the evolving landscape of digital education tools [182].

The field of HCI emerged in the 1980s alongside the rise of personal computing, with devices such as the Apple Macintosh, IBM PC 5150, and Commodore 64 becoming widely accessible. Previously, computers were large, costly machines used primarily by experts in specialised environments. However, as computing technology became more common in homes and workplaces, there was an urgent need to develop interfaces that were both functional and intuitive for general users. This shift led to the development of HCI as a multidisciplinary domain, incorporating elements of computer science, cognitive science, and human-factors engineering [182].

By integrating HCI principles into our research, we aimed to ensure that digital marking tools were not only efficient but also user-friendly for educators. The goal was to refine the interaction between teachers and assessment technology, making the marking process more intuitive and accessible. As technology continues to evolve, the role of HCI remains critical in designing educational tools that enhance, rather than hinder, the user experience [182].

HCI soon became the subject of intense academic investigation [183]. Those who studied and worked in HCI saw it as a crucial instrument to popularise the idea that the interaction between a computer and the user should resemble a human-to-human, open-ended dialogue. Initially, HCI researchers focused on improving the usability of desktop computers (i.e., practitioners concentrated on how easy computers are to learn and use) [184]. However, with the rise of technologies such as the Internet and smartphones, computer use is increasingly moving away from the desktop to embrace the mobile world. Also, HCI has steadily encompassed more fields [182].

As HCI research evolved, it expanded beyond usability to consider broader aspects of user interaction, including accessibility, efficiency, and emotional engagement. Researchers began exploring how different user groups interact with technology, leading to the development of specialised interfaces for diverse populations, such as those with disabilities [183]. Additionally, the integration of artificial intelligence (AI) into HCI has enabled more adaptive and personalised experiences, allowing systems to respond dynamically to user behaviour. This shift highlights the growing importance of human-centred design in ensuring technology remains intuitive and effective for a wide range of users [185].

HCI looks to consider how humans and technology interact with each other [182]. While some approaches might look within a retrospective angle about how and why users interact with technology in a certain way, a big focus has been made in recent times on having a human-centred approach by having the design, building of the technology and testing all be revolved around having humans in the loop [185, 186]. This shift towards participatory design ensures that end-users are actively involved in shaping technological solutions, leading to tools and systems that are better aligned with real-world needs [182].

The field of HCI and user-experience design (UX design) is very closely linked. Practitioners of HCI tend to be more academically focused. They're involved in scientific research and developing empirical understandings of users. Conversely, UX designers are almost invariably industry-focused and involved in building products or services, such as smartphone apps and websites. Despite this distinction, both fields share a common goal: improving the interaction between humans and digital systems. As technology continues to evolve, collaboration between HCI researchers and UX professionals remains essential in ensuring that emerging innovations prioritise usability, accessibility, and overall user satisfaction.

2.3.1 Semi-Structured Qualitative Studies

Semi-structured qualitative studies (SSQS) occupy a methodological space between ethnography and surveys, typically involving data collection techniques such as interviews and observations [187, 188]. These approaches are characterised by their flexible yet guided structure, allowing researchers to systematically code and analyse verbal data, often supplemented by other modalities. This balance enables the capture of rich, contextual insights while maintaining a degree of comparability across participants [189].

It is emphasised that the design of an SSQS should be purpose-driven [190], with methods tailored to address specific research objectives and practical considerations. It is suggested that there is not a one-size-fits-all approach, suggesting that the selection of data gathering and analysis techniques should be informed by the study's goals and constraints [191]. This perspective encourages researchers to thoughtfully integrate various qualitative methods to construct a coherent and effective research design.

The importance of systematic coding in SSQS and the development of themes that accurately reflect the data is important [189]. However, there are challenges around reporting qualitative findings, advocating for transparency in methodological decisions and clarity in presenting results to ensure the credibility and utility of the research [192].

In conclusion, qualitative research methods, including semi-structured approaches, have long been valued for their ability to generate theory grounded in empirical data [193]. By thoughtfully designing studies that align with specific research purposes and constraints, and by employing flexible yet systematic methods, researchers can gain deep insights into user experiences and interactions with technology [194]. This approach not only enriches the understanding of user needs but also informs the development of more effective and user-centred interactive systems.

2.4 Evaluation Methodology

2.4.1 Datasets

We utilise three real-world datasets: the DREsS corpus of English-as-a-Foreign-Language (EFL) essays [195], an undergraduate assignment dataset from a British university (BU), and a taught postgraduate (UK Level 7) dataset of critical reviews (PG). Each dataset comprises item identifiers and absolute marks assigned by human assessors. These marks serve to construct ground-truth target ranks and to drive the simulation of decision-making processes (Section 2.4.3). The datasets differ in rubric structure and weightings, enabling evaluation of the proposed active ranking strategies across heterogeneous assessment contexts.

2.4.1.1 DREsS (EFL essays)

The DREsS dataset contains $\sim 1{,}700$ essays produced by undergraduate EFL learners in authentic classroom settings [195]. It includes three sub-datasets: $DREsS_{\text{New}}$, $DREsS_{\text{Std.}}$,

and $DREsS_{\text{CASE}}$, each targeting different automated essay scoring (AES) scenarios. We focus on $DREsS_{\text{New}}$, which best reflects contemporary EFL writing conditions. Trained English education experts score essays on three equally weighted criteria—*Content*, *Organisation*, and *Language*—each on a 0–5 scale, yielding totals out of 15 [195].

2.4.1.2 BU undergraduate web assignment

The BU dataset comprises 69 submissions from first-year students. Students selected one scenario and developed a web page demonstrating specified skills. For consistency, we analyse a subset of 38 items based on the same scenario. Scripts were assessed on three criteria: implementation quality of core components, fulfilment of additional brief requirements, and documentation quality. Each criterion was scored on a 0–100 scale and combined using weightings of 50%, 25%, and 25%, respectively, to produce the final mark.

2.4.1.3 Postgraduate critical review (PG) dataset

We obtained marks for 30 submissions from a Level 7 taught-postgraduate course in which students wrote a $\sim 1{,}000$-word critical review of a recent research paper. Submissions were anonymised (IDs unlinked to student identities) by the lead lecturer, hereafter the "Oracle". For downstream comparisons of marking approaches, we created three groups of 10 items each for traditional rubric-based marking, BCJ, and MBCJ, using stratified sampling [196]. The official, ratified Oracle marks (out of 20) define the target rank; the distribution of marks is shown in Figure 5.1. The assessment rubric covered five areas: *Introduction and Summary*, *Quality of Analysis and Evaluation*, *Conclusions*, *Writing Quality*, and *References*; the assignment brief was provided to markers in advance.

2.4.1.4 Experimental subsets and budgets.

For controlled experiments, we draw stratified subsamples of size $N \in \{5, 10, 15, 20, 25\}$ from each dataset to ensure coverage across the mark range [196]. Given a subsample of size N, we set a comparison budget via a multiplier $K \in \{5, 10, 20, 30\}$, leading to $N \times K$ pairwise judgements under each strategy. Ground-truth target ranks are derived from the absolute marks within each subsample.

2.4.2 Metrics

Evaluating the performance of ranking methods requires metrics that capture both rank accuracy and distributional fidelity.

2.4.2.1 Rank Accuracy

To quantify rank accuracy, we use the normalised Kendall's τ rank distance, which measures the proportion of discordant pairs between two rankings π and σ [197, 198]. Formally, for n items,

$$d_\tau(\pi, \sigma) \;=\; \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \mathbf{1}\Big[\,\mathrm{sgn}\big(\pi(i) - \pi(j)\big) \neq \mathrm{sgn}\big(\sigma(i) - \sigma(j)\big)\Big],$$

so $d_\tau \in [0, 1]$, where 0 indicates perfect agreement and 1 indicates complete disagreement. When there are no ties, Kendall's correlation and the distance are related by $\tau \;=\; 1 - 2\,d_\tau$, making d_τ directly interpretable as the fraction of pairwise orderings that differ (e.g. $d_\tau = 0.03$ means 3% of pairs are discordant). In the presence of ties, tie-aware variants (e.g. τ_b) can be used with the corresponding normalisation.

2.4.2.2 Distributional Accuracy

To evaluate how closely the estimated score distributions match the ground truth, we use the Jensen–Shannon Divergence (JSD), a symmetric and bounded measure derived from Kullback–Leibler divergence [199, 200]. Given two probability distributions P and Q, JSD computes their average $M = \frac{1}{2}(P + Q)$ and then calculates:

$$\mathrm{JSD}(P\|Q) = \frac{1}{2}\mathrm{KL}(P\|M) + \frac{1}{2}\mathrm{KL}(Q\|M),$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence [199, 200]. JSD values range from 0 (identical distributions) to 1 (maximal divergence, using $\log_2$), making it an interpretable metric for distributional similarity. In our context, JSD quantifies how well BCJ captures the full uncertainty structure of item scores compared to the target distributions [200].

2.4.2.3 Statistical Comparison

Each strategy is run independently for a fixed budget over 50 trials. We compare final τ scores using the one-tailed Wilcoxon rank-sum test (Mann–Whitney U) [201], applying

a Bonferroni correction [202] for multiple comparisons: $\alpha_{\text{adj}} = 0.05/S$, where S is the number of strategies. Additionally, we compute $V(i)$, the count of times method i is significantly outperformed by others:

$$V(i) = \sum_{i \neq j \wedge j \in S} \big[\text{p-value}(i > j) \leq \alpha_{\text{adj}}\big].$$

2.4.3 Automated Decision Simulation

In the absence of human decision-makers — although we conduct real experiments involving human participants in chapter 5 — it is necessary to simulate the decision-making process in a plausible and principled way. To achieve this, we adopt the method where each item's ground truth mark is treated as the mean of a Normal distribution. The standard deviation reflects typical tolerance levels observed in marker disagreements.

Regarding the DREsS dataset, where no formal guidance on tolerance is available, we set the standard deviation to $\sigma = 0.5$. This corresponds to a discrepancy of $\pm 2\sigma = \pm 20\%$ in absolute marks, encompassing approximately 95% of the probability mass under a Normal distribution. This represents a relatively relaxed marking scenario. In contrast, for the BU dataset, we follow established tolerance guidelines and set the standard deviation to $\pm 3\%$, resulting in an acceptable discrepancy of approximately $\pm 6\%$ with 95% confidence. These two contrasting scenarios allow us to explore the impact of uncertainty in the marking process — from high variability in DREsS to a more stringent and consistent marking regime in BU.

Under this setup, we simulate pairwise comparisons by sampling from two Normal distributions. Each item's score is drawn randomly, and the item with the higher score is considered the winner. Formally, for items i and j, we sample $x_i \sim \mathcal{N}(\mu_i, \sigma_i)$ and $x_j \sim \mathcal{N}(\mu_j, \sigma_j)$, where $\mathcal{N}(\mu, \sigma)$ denotes a Normal distribution with mean μ and standard deviation σ. The simulated winner is determined as follows:

$$w_{i,j} = \begin{cases} 1 & \text{if } x_i \geq x_j \\ 0 & \text{otherwise} \end{cases} \tag{2.22}$$

2.4.3.1 An Illustration

We consider a set of *five* items with respective scores and the associated uncertainties, as shown in Figure 2.6. We assume that the scores are Normally distributed (as per

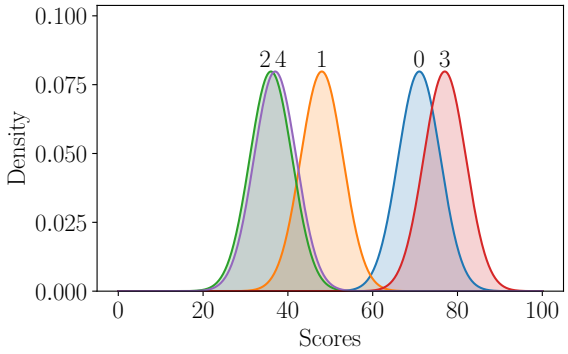


Figure 2.6: An illustration of five items with Normally distributed scores. Here, the mean vector for the items is $\boldsymbol{\mu} = (71, 48, 36, 77, 37)^\top$, and $\sigma = 5$ represents the uncertainty around the mean scores. The σ also represents the range of marks multiple judges could give the piece of work from absolute marking that would still result in the work being within tolerance level, which in this case is a 10 mark tolerance on either side of the given mark, therefore meaning that there is a 95% chance that the difference between the markers would be 10 or less. A simulated paired comparison entails sampling from a pair of these distributions, and the distribution that yields the higher score wins.

Thurstone's original work). To generate the means of these distributions, we uniformly sampled N numbers between 30 and 90. Typically, it is often acceptable to have ± 10 score difference between markers when the scores are on a scale between [0,100]. So, we set the two standard deviations of the distributions to 10, i.e. $2\sigma = 10$. It should be noted that these assumptions about score ranges and standard deviations are only for illustration purposes. The method presented in this chapter does not rely on these and can work with arbitrary distributions over the scores.

With Normal distributions over scores, we can compute the probability distributions over ranks for any item using the formula in (3.11) as we can calculate the probability that i dominates j as follows [203]:

$$P(i \succ j) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{m}{\sqrt{2}}\right)\right], \tag{2.23}$$

with $m = \frac{\mu_i - \mu_j}{\sqrt{\sigma_i^2 + \sigma_j^2}}$ where μ_i and μ_j and means of the Normal distributions for i and j, and the associated standard deviations are σ_i and σ_j. The function $\text{erf}(\cdot)$ represents the Gauss error function [204].

2.5 Reflecting on Prior Work

One of the main limitations of CJ is its perceived opacity in how rankings are generated. Unlike traditional rubric-based marking, CJ does not provide explicit justifications for why one submission is ranked higher than another, making it difficult for educators, students, and stakeholders to understand the rationale behind assessment outcomes. This lack of transparency raises concerns about fairness and trust in the system, particularly in high-stakes assessments. Addressing this issue requires the development of a more interpretable CJ framework that clarifies ranking decisions, potentially through the integration of confidence intervals, justification metrics, or visualisation techniques that illustrate the decision-making process.

Another critical gap in CJ research is the challenge of translating comparative rankings into meaningful grades. Traditional assessments use absolute grading criteria, allowing students to receive individualised scores based on performance thresholds. In contrast, CJ primarily produces a rank order, which does not inherently map to standardised grading systems. The absence of a direct conversion method makes it difficult to align CJ outcomes with established grading practices. Developing a systematic and fair approach to convert CJ ranks into grades—possibly using statistical normalisation techniques or probabilistic modelling—would make CJ more practical for widespread adoption in educational settings.

Traditional assessment often involves evaluating student work across multiple dimensions, such as clarity, originality, and technical accuracy. However, standard CJ methods collapse these criteria into a single holistic judgement, potentially losing valuable information from rubrics. Multi-dimensional CJ seeks to retain this structure by enabling assessors to compare work across distinct learning outcomes independently, ensuring a more granular and informative assessment process. Implementing this approach requires the development of CJ models that can handle multi-criteria evaluations while maintaining efficiency and reliability in ranking generation.

Beyond improving the transparency of decision-making in CJ, another critical challenge is how rankings are presented to educators and students. Current CJ implementations often provide only a final rank order, with limited insight into the degree of certainty or differentiation between rankings. More effective ways of demonstrating ranks could include visual analytics, confidence bands for ranking, or interactive tools that allow users

to explore the rationale behind rankings. These approaches would not only increase stakeholder trust in CJ but also support students in better understanding their performance relative to their peers.

A significant concern in CJ is the variability in assessors' decision-making, which can affect the reliability of rankings. Traditionally, Scale Separation Reliability (SSR) has been used to measure the consistency of judgements within CJ. However, SSR has limitations in capturing nuanced patterns of assessor agreement and disagreement. Alternative methods, such as Bayesian reliability modelling or network-based ranking algorithms, could provide a more refined evaluation of assessor performance, improving the robustness and fairness of CJ rankings. Developing alternative rating models for assessor comparisons would enhance CJ's credibility as a fair and objective assessment method.

With our novel approach, we aim to use Bayes from the ground up. Fundamentally rewriting the process that most methods of CJ use. We are using Bayes to inform us at every step, rather than conducting comparisons. Once completed, we compile the exact comparisons and results into a BTM and apply our Bayes approach to examine the results generated from the comparisons and assess their performance.

This review has explored the theoretical and practical landscape of educational assessment, highlighting the limitations of traditional methods and the emergence of CJ as a promising alternative. However, challenges around efficiency, bias, and transparency remain unresolved. These gaps motivate the research presented in this thesis: to develop Bayesian and ML-based methods that enhance the scalability, interpretability, and educational usefulness of comparative judgement systems.

The literature review highlighted both the promise and limitations of existing CJ approaches, particularly in terms of transparency, efficiency, and reliability. Addressing these challenges requires innovative methods that retain the strengths of CJ while overcoming its limitations. Building on these insights, the next chapter introduces BCJ, a novel framework that leverages probabilistic modelling and active learning to improve accuracy, efficiency, and transparency in assessment. This chapter outlines the theoretical foundations of BCJ and demonstrates its effectiveness through both synthetic and real-world experiments.

Chapter 3

Bayesian Comparative Judgement for Holistic Pair-wise Comparisons

Building on the limitations and opportunities identified in the previous chapter, this chapter develops and presents a novel BCJ to strengthen the CJ process. By combining Bayesian inference with active learning, BCJ addresses concerns of ranking reliability, efficiency, and transparency while modelling uncertainty in a way that traditional CJ methods cannot. The chapter outlines its theoretical foundations, implementation, and initial experimental validation, providing the basis for the subsequent extension to multi-criteria approaches, which will be explored in the following chapter.

The core mathematical technique used for generating ranks from paired comparisons in comparative judgement for assessment was proposed in 1927 [135]. In this chapter, for the first time, we propose a Bayesian approach that appropriately considers the epistemic uncertainty arising from a limited number of comparisons and propagates it to estimate predictive uncertainty in the derived ranks while coping with the aleatoric uncertainty in judgements from a single or multiple assessors. This enables the assessor to make an informed decision on the ranks and grades of submissions under uncertainty. We expect this to bring about a paradigm shift in the way comparative judgment is conducted for assessment in practice.

A key limitation of CJ is the considerable time required for marking, collating grades, awarding scores, and providing feedback. ACJ seeks to reduce the number of comparisons without compromising accuracy, but can introduce other biases due to its adaptive nature [148], making the pure form of CJ still the preferred choice. This underscores an

ongoing research challenge: developing a method that preserves CJ's advantages while reducing both comparisons and marking time. Additionally, Ofqual notes that CJ's rank order can deteriorate unless the minimum number of judgments is precisely determined, which is currently unknown [147], and highlights concerns regarding the transparency of how CJ produces and presents results [147].

We thus propose a novel Bayesian approach to CJ, which we name BCJ, addressing the key weaknesses of traditional CJ. Our primary aims in developing BCJ were to reduce interactions and provide greater insight into the ranking decision process. The main contributions of this chapter are as follows:

- We derived an analytical expression to compute the entire *predictive rank distribution* for any item that is being assessed with densities over pairwise preferences.
- We illustrate how each of these pairwise preference densities and, as a consequence, the overall rank distributions for an item, can be updated via Bayesian methodology, as we collect more data on pairwise comparisons.
- We propose a novel active learning (AL) approach, based on predictive entropy of the pairwise preference densities, i.e. a measure of the average uncertainty about the outcome of the contest, to select the next pair that should be assessed.
- We propose a probabilistic approach based on predictive rank distributions to assign a grade to each item, in a norm-referenced manner, controlled by the assessor.
- For the first time, we demonstrate through repeated experiments on a range of synthetic problems that the proposed BCJ AL framework with an entropy-based selection method is statistically the best (or equivalent to the best, i.e. the most accurate in estimating a ground truth rank in the presence of uncertainty) for all configurations.
- We demonstrate BCJ in a real dataset from [156], and highlight the advantages of the proposed method in comparison to standard CJ.
- We introduce two novel metrics – Mode Agreement Percentage (MAP) and Expected Agreement Percentage (EAP) – based on a Beta prior over pairwise preferences. These metrics quantify the level of agreement among assessors and help identify

controversial comparisons. In particular, EAP serves as a direct indicator of reliability under uncertainty due to limited data and offers a principled stopping criterion for data collection.

The rest of the chapter is structured as follows: section 3.1 outlines how the main algorithms work to rank students' work. We will explain the new novel method for selecting the next pairs to be compared in Section 3.2. We present our results and discussions in Section 3.3, followed by general conclusions and future work in Section 4.5.

3.1 Bayesian Comparative Judgement

To enable us to use a Bayesian approach for our CJ system, we implied that every outcome of a paired comparison is a Bernoulli-distributed outcome. This is because the Bernoulli distributions are used when dealing with binary values, such as a win (1) or a loss (0). In probability theory, the Bernoulli distribution is a discrete probability distribution that describes the outcome of a single binary event. It is a special case of the binomial distribution and is parameterised by a single probability π, representing the probability of the event occurring [205].

Ultimately, we treat each paired item combination as a coin flip. Therefore, we want to find its bias in its flips or, in the concept of our CJ application, the bias towards what item in each pair. For our application, this relates to the probability that item a will beat item b, with the bias representing that value [205]. Using a grid approach, we can see the maximum likelihood value for each distribution after every comparison.

As established, the Bernoulli distribution is often used to model events with two possible outcomes: a coin flip or a Boolean value (e.g. true or false). It has many properties, one being X representing a random variable. Therefore, we can represent the distribution as follows [205]:

$$Pr(X = 1) = p = 1 - Pr(X = 0) = 1 - q \tag{3.1}$$

In equation 3.1, we can see that the probability of $X = 1$ equals the probability, which is one minus the probability of $X = 0$, which can be referred to as q, which then in itself can be referenced as $1 - q$ [175].

Once we establish possible outcomes, we can use the probability mass function (PMF). Therefore, as long as the event has a probability p of occurring, we can find the PMF of the Bernoulli distribution by [205]:

$$f(k;p) = \begin{cases} p & \text{if } k = 1, \\ q = 1 - p & \text{if } k = 0 \end{cases} \tag{3.2}$$

We use equations 3.1 and 3.2 to calculate our **posterior conjugate** for our Bayesian calculations. Therefore, it is the crucial first step in determining our items' final ranks.

While the current CJ based on BTM works well, a core weakness is that it produces a point estimate of performance by maximising the likelihood in equation 2.5 without estimating the epistemic uncertainty in ranks due to the paucity of data. One way to estimate the uncertainty (that is not commonly used in an education context) is to use a Bayesian statistical approach; interested readers should refer to van de Schoot *et al.* [206] for a concise and recent overview, and to McElreath

[207], or Lambert[208], for a complete and accessible discourse on the topic.

Typically, the application of a Bayesian approach to CJ has involved using *prior distributions* over the performance vector $\boldsymbol{\gamma}$ (and other parameters of the likelihood function) alongside the observed data to identify a posterior distribution over $\boldsymbol{\gamma}$ using Bayes' theorem [209, 169], and produces similar results to standard CJ in terms of identifying the ranking [170, 171]. However, there are important barriers that make it challenging to adopt for real-world deployment. Two key issues are:

Computation Time. *Inferring* the posterior distribution requires computationally expensive sampling-based approaches (e.g Markov Chain Monte Carlo [170]), as an analytical solution to computing the posterior is usually not available in this context. This is a major issue in using this approach for practical implementations: we want to be able to indicate the ranks to the assessors quickly, possibly after each pairwise comparison, without a significant delay (e.g. several minutes).

Modelling Performance Instead of Pairwise Preference. In a ranking exercise, we are generally interested in identifying the ranks of the items, and the observed data is from pairwise comparisons. However, in standard CJ, including the typical Bayesian approach, the performances are modelled instead of pairwise preference; the latter

is usually treated as an outcome of a latent function and thus only reflected in derived ranks from the expected (or average) performances. As a result, while it is possible to extract uncertainty estimates over the preferences or the ranks (with the aforementioned computational expense), they are never communicated or used to provide insights to the assessors. Subsequently, an opportunity to utilise the uncertainty in preference to drive the collection of new pairwise comparisons is missed. Furthermore, the performance scores that result from these models do not have a direct scalar relationship to the scores of the assessment designed by the assessor. Therefore, it is difficult to easily interpret these scores.

Addressing these primary issues, we propose to adopt a Bayesian approach where we focus on *modelling pairwise preferences*. We expect that this approach will allow us to capture most information because of the direct relationship between pairwise preference and data from pairwise comparisons. The posterior allows us to identify the predictive density over the ranks of the items. Moreover, the uncertainty estimations in preferences help us drive the selection of the next pair to compare in an active learning manner. We discuss the selection method in Section 3.2.

3.1.1 Pairwise Preference Model

Let the result of a paired comparison between the ith and jth item be binary, i.e. $x = 0$, or $x = 1$, with $x = 1$ representing a preference for i and *vice versa*. Now, considering the data $\mathbf{x} = (x_1, \ldots, x_n)^\top$ as results of n comparisons, we can calculate the number of wins $w = \sum_{k=1}^{n} x_k$. With these results of the Bernoulli process, the likelihood can be defined as [168]:

$$L(p|\mathbf{x}) \propto p^w (1-p)^{n-w}. \tag{3.3}$$

In Bayesian probability theory, for certain likelihood functions, there exists a conjugate prior, where the prior and posterior are in the same family of distributions. This enables fast and analytical computation of the posterior. For the likelihood above, the conjugate prior is known to be a Beta distribution with two shape parameters $\alpha > 0$ and $\beta > 0$. The posterior Beta density $\pi(p|\mathbf{x}, \alpha_{init}, \beta_{init})$ simply uses the following rule for updates [210]:

$$\alpha \leftarrow \alpha_{init} + w, \tag{3.4}$$

$$\beta \leftarrow \beta_{init} + (n - w). \tag{3.5}$$

With priors of $\alpha_{init} = 1$ and $\beta_{init} = 1$, we obtain a uniform prior, meaning we do not have any prior preference between items at the beginning of the CJ process. Henceforth, for notational simplicity, we remove $\mathbf{x}$, α_{init} and β_{init} from the equations. As we collect data, the density changes its shape through the updates in α and β; an example is given in Figure 3.1. Clearly, this update can be done as a sequential process or all together at the end of the data collection, and it can be rapidly performed for a pair for any amount of data.

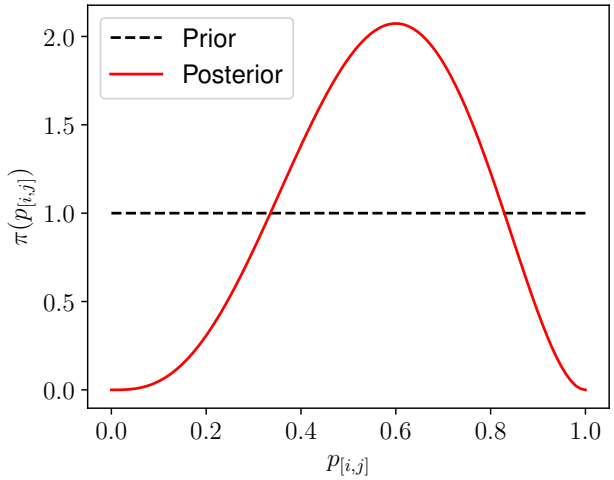


Figure 3.1: A toy example of Bayesian updating of PDF over preference between ith and jth items. Initially, with a uniform prior (shown with a black dashed line), none is preferred. Then, with three wins ($\alpha = 1 + 3 = 4$) and two losses ($\beta = 1 + 5 - 3 = 3$) for i after five comparisons, the PDF (depicted with a red line) starts to skew in favour of i (i.e. towards 1). The more data we have, the narrower the PDF will become, i.e. the uncertainty will reduce.

With this framework, we define the probability that i is preferred over j, i.e. a different interpretation of probability of winning in (2.4), as:

$$P(i \succ j) = P\left(\pi(p_{[i,j]}) > 0.5\right) = 1 - \mathcal{F}(0.5), \tag{3.6}$$

where $\mathcal{F}(\cdot)$ is the cumulative distribution function (CDF) for the Beta PDF $\pi(p_{[j,i]})$. Using symmetry, we can calculate the probability that j will be preferred over i as:

$$P(j \succ i) = 1 - P(i \succ j). \tag{3.7}$$

We now extend this analysis to the N items and discuss the computation of the distribution over ranks based on this model.

3.1.2 Distribution Over the Rank of an Item

For a set of N items, we therefore define a $N \times N$ matrix $\mathcal{P}$, where each cell holds a PDF $\mathcal{P}_{[i,j]} = \pi(p_{[i,j]}) \mid i \neq j$ defined by a respective $\alpha_{[i,j]}$ and $\beta_{[i,j]}$ updated in a Bayesian manner based on observed data. The diagonal of this matrix is essentially empty, as it does not make sense to construct a preference density for the same item paired with itself. Now, due to the symmetry discussed in (3.7), we are only required to consider the upper triangle of this matrix for updates, which is fast to compute, even for large N.

The ith row $\mathcal{P}_{[i,:]}$ captures the relationship between i and other components in the set I. Now, to compute the probability that an item is ranked at the top, we must consider all the constituent probabilities that the item dominates each of the other individual items. To be precise, it must simultaneously dominate all other items in the set of all items; hence, this aggregation should be done with the product rule assuming independence between the preferences for the ith item when compared with each of the other unique items. We can write down the expression for computing this probability as follows (with 1 being the top rank):

$$P(r_i = 1) = \prod_{j \in I \setminus \{i\}} P(i \succ j). \tag{3.8}$$

Similarly, we can compute the probability that an item is ranked at the bottom as:

$$P(r_i = N) = \prod_{j \in I \setminus \{i\}} P(j \succ i). \tag{3.9}$$

For generalisation, specifically for intermediary ranks, for an arbitrary rank a, first consider a set $O = I \setminus \{i\}$ with cardinality $|O| = N - 1$. Now, for i to be in rank a, there must be $a - 1$ *dominant* items. From set O, we can pick $z_a = C_{N-1,a-1} = \frac{(N-1)!}{(N-a)!(a-1)!}$ combinations without repetitions that can be considered as dominating the ith item. For every kth combination, we then split O into two sets: one for dominant items D_k and the other for dominated items E_k, where $|D_k| = a - 1$, and $|D_k| + |E_k| = |O|$. For kth combination with D_k and E_k, the component probability that i is ranked a is:

$$P(r_i = a | D_k, E_k) = \prod_{s \in D_k} P(s \succ i) \prod_{t \in E_k} P(i \succ t). \tag{3.10}$$

Expanding on this, the total probability that i is ranked a can be expressed as:

$$P(r_i = a) = \sum_{k=1}^{z_a} P(r_i = a | D_k, E_k), \tag{3.11}$$

which for a range of $a \in [1, N] \subset \mathbb{N}$ is a discrete probability distribution, and adheres to the property $\sum_a P(r_i = a) = 1$. The expected (i.e. average or the first moment) rank of an item i can thus be computed using:

$$\mathbb{E}[r_i] = \sum_a aP(r_i = a). \tag{3.12}$$

Now, the number of component combinations that construct the complete probability density for an item is $\sum_{l=1}^{N} z_l$. Thus, to repeat the procedure for all items, it would require $N \sum_{l=1}^{N} z_l$ components to be identified and computed. For example, with 25 items, there will be over 419 million components. While each component is fast to compute, with a large number of components, it may be computationally expensive to compute the complete probability density for all items.

A straightforward way to combat the expense of computing the expected rank of an item in (3.12) is to use a form of numerical integration. In fact, a simple Monte Carlo (MC) integration [211] with a large *enough* number of samples would be effective in this case (as we illustrate in the next section). To perform MC estimation of the expected rank of an item i, we first take samples from the respective row of the matrix $\mathcal{P}$: this generates a sample vector $\mathbf{x}'_i = (x'_{[i,j]})^{\top}_{j \in [1,N] \wedge i \neq j}$, where $x'_{[i,j]} = \lfloor X \rceil \mid X \sim \mathcal{P}_{[i,j]}$. This allows us to count the number of times i has won a comparison $w' = \sum_{j \in [1,N] \wedge i \neq j} x'_{[i,j]}$. Naturally, the rank is $r'_i = (N + 1) - w'$; c.f. with (2.8). For R samples, we can then estimate the expected rank of i as follows:

$$\mathbb{E}[r_i] = \frac{1}{R} \sum_{k=1}^{R} r'_i[k], \tag{3.13}$$

where $r'_i[k]$ is the kth sampled rank for i.

The standard error of this estimate is known to be $\frac{\sigma_s}{\sqrt{R}}$ with σ_s as the standard deviation of the samples [212]. In other words, the standard error reduces at the rate of $\frac{1}{\sqrt{R}}$. It is typical to use 10k samples for this approximation method. In this case, we would need $10,000$ samples to estimate ranks for all items, which can be done efficiently on a standard desktop computer, even for large N.

To determine the final rank of the items, we sort items by their expected ranks:

$$r_{i \in I} = (N + 1) - \text{argsort}(\mathbb{E}[\mathbf{r}]). \tag{3.14}$$

We present an illustrative synthetic example in the following section as explained in chapter 2.4.3.1.

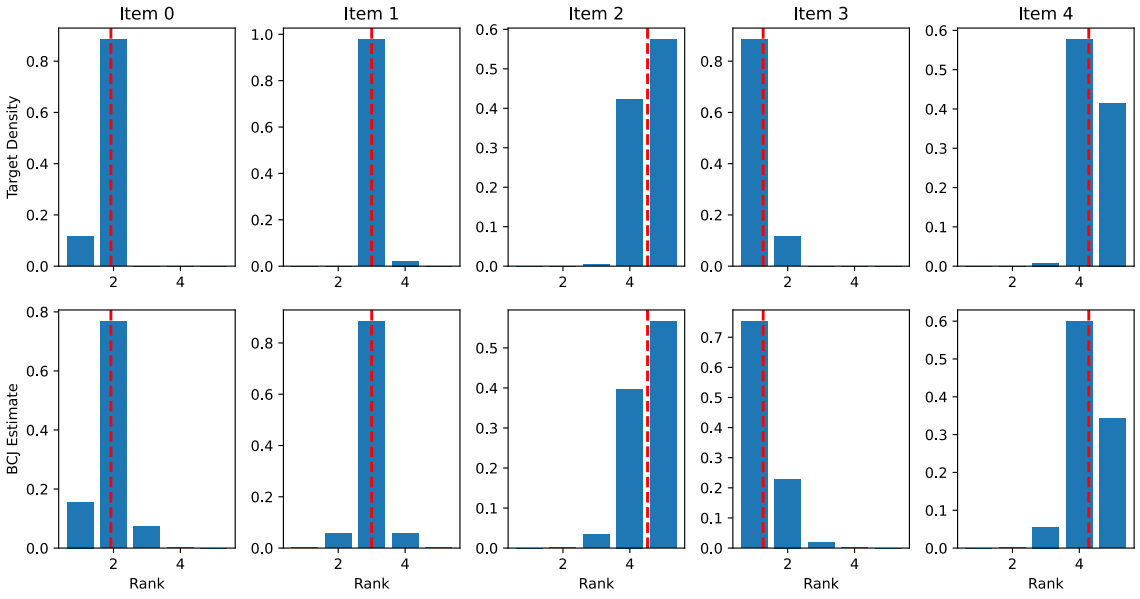


Figure 3.2: Probability distributions of ranks of items presented in Figure 2.6. The top row shows the densities calculated directly from the Normal distributions over the scores using (3.11). The bottom row shows the estimated rank distributions using our proposed BCJ method after $N \times K = 5 \times 10 = 50$ pairwise comparisons (driven by our entropy-based active learning method presented in Section 3.2). The red dashed vertical line in each panel depicts the expected rank for relevant density. Clearly, our method can accurately estimate the target densities, as well as the expected rank vector $\mathbb{E}[\mathbf{r}]$.

3.1.3 A Ground Truth Comparison Illustration

In Figure 3.2, we show the target distribution over ranks for the items in Figure 2.6, calculated using (3.11) and (2.23). In this case, to emulate the result of a comparison, we sample from the pair of densities, and whichever produces a higher score wins the duel. After completing $N \times K = 5 \times 10 = 50$ comparisons using our proposed BCJ method, we can easily approximate the target distributions. To measure how close the estimated distribution is, we use the Jensen-Shannon divergence (JSD). This measure is based on the Kullback–Leibler divergence, with some notable differences, including that it is symmetric and always has a finite value between 0 and 1 [213] with values 0 representing a perfect match. In this case, we get the JSD values of $0.0299, 0.0254, 0.008, 0.0185$, and 0.0125, which are reasonably close to 0.

It should be noted that with the traditional BTM-based CJ, we cannot obtain an estimate of the probability densities over the ranks, and hence, it is impossible to compute an average rank in this manner. In that method, the scores are used to rank the items

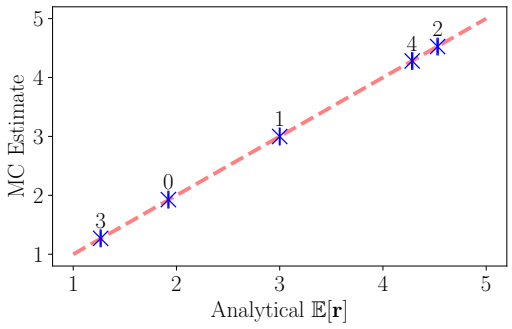


Figure 3.3: Comparison between the analytical estimates in (3.11) and Monte Carlo estimates (with 10k samples) in (3.13) of the expected rank vector of the items $\mathbb{E}[\mathbf{r}]$ for our proposed BCJ method after $N \times K = 5 \times 10 = 50$ comparisons as in Figure 3.2. Crosses show the mean MC estimate, and the vertical error bars represent the respective uncertainty in approximation, and, as expected, they are reasonably small for the 10k samples. The red dashed line shows when there is perfect agreement between the analytical and estimated values, and we see that the average MC estimates are (almost) perfect.

instead. To compare our approach with BTM-based CJ, we will therefore use the BCJ expected ranks to identify the ranks of the items.

In Figure 3.3, we show a comparison between analytical and MC estimates of rank distributions of items with the BCJ process. Clearly, the MC estimates are highly reliable. So, for large N, we recommend using MC estimates to generate expected ranks. In this chapter, we use the analytical approach from now on.

In the next section, we discuss the selection of pairs to evaluate the problem and relevant solutions, including our entropy-driven approach.

3.2 Active Learning

Within our experiment, we applied active learning (AL) to determine which pairs to select for our comparisons that would provide the most information. We are then using Bayesian optimisation's entropy-based approach to calculate the uncertainty values for the pairs to determine which pairs we are the most uncertain about (see section 2.2.3.3.

Figure 3.4 demonstrates the entropy score after each round of comparisons, which can then be used as the selection process over the n required number of rounds. The process involves the algorithm calculating the entropy value for each pair combination to see which pair has the highest value and then selecting that pair to be presented. However,

suppose there are multiple combinations at the same entropy score. In that case, the algorithm will randomly select a pair of values from the list of combinations with the same entropy value. This process will repeat until the required number of rounds is reached.

We have developed a novel approach to selecting pairs in the context of CJ, which uses a Bayesian active learning (AL) approach. AL is a subcategory of machine learning in which a learning algorithm can request input or labels from a user or any other source of information to label new data points [214, 215, 216]. In Bayesian AL, we use a Bayesian model to make predictions and then actively select the next data points that should be labelled via an acquisition function that identifies the utility of augmenting the dataset with this new data point; see, for instance [217]. In this way, we collect data efficiently and learn a good model with fewer data points.

There are many variants of AL. In this chapter, we focus on so-called "pool-based learning" [218] where we have a finite set of options, and we are going to choose one to show to the labeller. The simplest acquisition function in this context is known as *uncertainty sampling*, where the option with the highest uncertainty is selected for labelling [219].

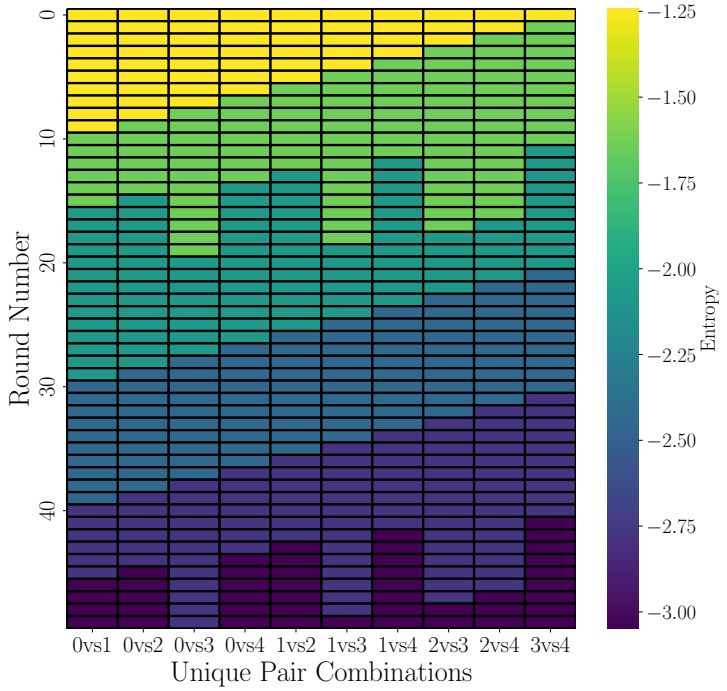
To be precise, in our context, we have a finite set of pairs of items, and we will select the one with the highest posterior uncertainty. This uncertainty can be measured with *entropy*, where higher uncertainty is represented by higher entropy, and for the posterior Beta density of BCJ, it can be computed as [220]:

$$\begin{aligned} H\left[\pi(p_{[i,j]})\right] &= \ln B(\alpha,\beta) - (\alpha-1)\psi(\alpha) \\ &-(\beta-1)\psi(\beta) + (\alpha+\beta-2)\psi(\alpha+\beta), \end{aligned} \tag{3.15}$$

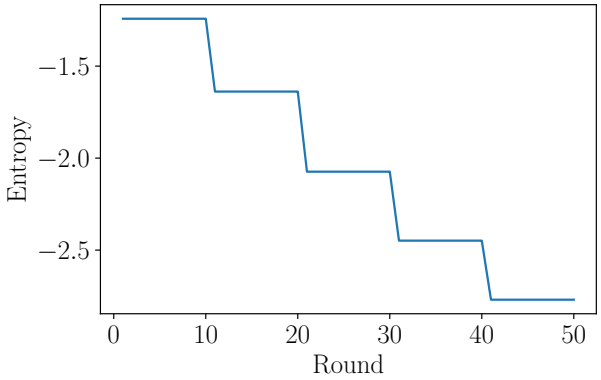
where, $B(\alpha,\beta)$ is the Beta function and the $\psi(\cdot)$ is the Digamma function. Given the parameters α and β, this calculation is straightforward using existing statistical packages, such as `scipy.stats` in Python [221].

In this chapter, we propose to locate the cell in the matrix $\mathcal{P}$ that has the highest entropy and select that pair to be presented to the assessor for making a choice on the preferred item.

In Figure 3.4, we demonstrate the entropy score after each round of comparisons, and the associated selection process. The process involves the algorithm calculating the entropy value for each pair combination in $\mathcal{P}$ to see which pair has the highest value, and



a) Entropy score for each unique combination after every pairing round. A higher entropy value shown in a lighter colour shows a higher uncertainty.



b) Progression of the highest entropy value after every AL round.

Figure 3.4: Illustration of uncertainty sampling using entropy (*top*) for the five items in Figure 2.6 after $N \times K = 50$ comparisons, and the respective gradual reduction in maximum entropy (*bottom*). As a pair is selected, its uncertainty immediately reduces after data is gathered about preference. The downward trajectory in maximum entropy shows that the model is becoming more accurate over iterations.

then selecting that pair to be presented. However, if there are multiple combinations at the same entropy score, the algorithm will randomly select a pair of values from the list of combinations with the same entropy value. This process repeats until the required number of rounds is reached. As we can see, the process may be similar to a round-robin approach, but our method would adapt to the changing uncertainties in the target densities in Figure 2.6.

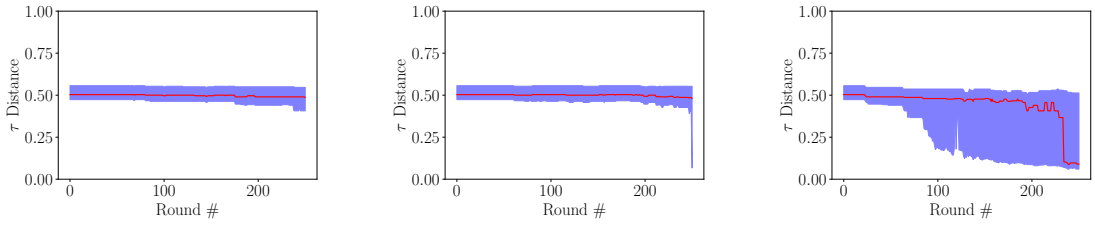
Entropy-based active learning not only improves computational efficiency but also introduces an interpretable rationale for pair selection. In practice, this could help educators understand why particular comparisons are made and where the system needs more information. This addresses known issues with traditional CJ systems, such as opaque decision logic and distrust in automated processes. By making pair selection adaptive and transparent, this approach makes CJ more accessible and trustworthy for real-world educational use.

3.3 Experiments and Discussions

In this section, we will present our findings, analyse them, and discuss what we believe they represent and mean.

In our reading of the literature, we found that the suggested budget for the number of comparisons was $N \times K = 10N$ [154]. However, in practice, a larger budget is often used. To identify what K allows different CJ methods to produce reasonable performance, we ran experiments with $K \in \{5, 10, 20, 30\}$.

As discussed, we have two rank generation methods: BTM and BCJ, and three pair selection methods: random (R), no repeating pairs (NR), and entropy (E) driven AL. Taking all possible combinations of rank generation and pair selection methods, we can construct a set of *six* approaches for CJ: $S = \{BTM^R, BTM^{NR}, BTM^E, BCJ^R, BCJ^{NR}, BCJ^E\}$. We run 50 repeated experiments for each approach in S for a given N and K, each time starting from scratch, to identify the best. These experiments were conducted with synthetically generated target distributions (following the methods elaborated in Section 2.4.3.1); these were paired, and therefore, we performed Wilcoxon Rank-Sum tests on the final results with Bon-Ferroni correction for multiple comparisons [222] at a significance level of $\alpha = 0.05$.



a) Random pairs selector with the BTM approach.

b) No repeating pairs selector with the BTM approach.

c) Entropy pairs selector with the BTM approach.

d) Random pairs selector with the Bayes approach.

e) No repeating pairs selector with the Bayes approach.

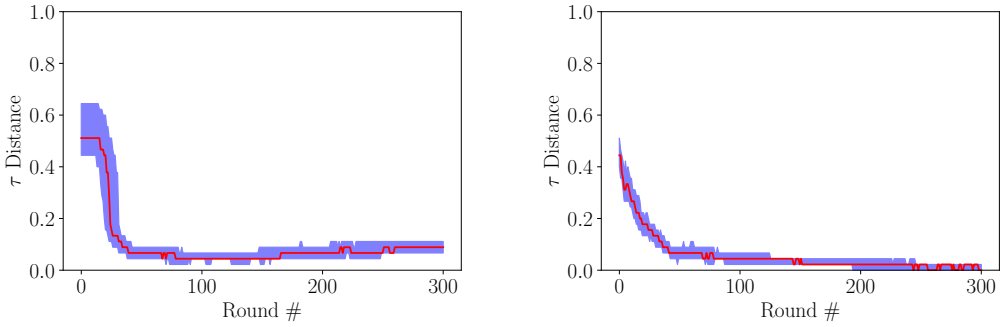
f) Entropy pairs selector with the Bayes approach.

Figure 3.5: A comparison of the random (3.5a), 3.5d)), no repeating pairs (3.5b), 3.5e)) and entropy (3.5c), 3.5f)) τ distance results. The light blue regions show performance between the 25^{th} and 75^{th} percent quartiles, and the red line depicts the median performance over 50 repetitions for 25 items where $K = 10$, making it a budget of 250 comparisons. The top row shows performances for BTM, while the bottom row shows respective results for our proposed Bayesian approach. Clearly, BCJ outperforms BTM throughout the progress towards the budget.

3.3.1 Analysing the Winning Method

In Figure 3.5, we first illustrate the convergence of each CJ approach for 25 items with a budget of 250 comparisons. We can see that overall, the BCJ approach has done better in all three pair selection methods. This is consistent across the board, with the BCJ and the novel entropy pair selection method generally being the best combination. The no-repeat selection method in combination with BCJ also performs well, but not as well as the combination of our two novel approaches. We also note that the entropy pair selection method positively impacts the BTM CJ approach.

To investigate Ofqual's claim that the performance of BTM-CJ with no repeating pairs deteriorates with many comparisons [147], we ran an experiment with $N = 10$ and $K = 30$ for both the current version of BTM-CJ with no repeating pairs and BCJ with entropy-based pair selection. Convergence plots are shown in Figure 3.6. The performance of BTM-CJ deteriorated over many iterations. However, it is difficult to determine the core reasons behind it. We suspect that this is because of the uncertainty in determining which



a) No Repeating pairs selector with the BTM approach.

b) Entropy pairs selector with the Bayes approach.

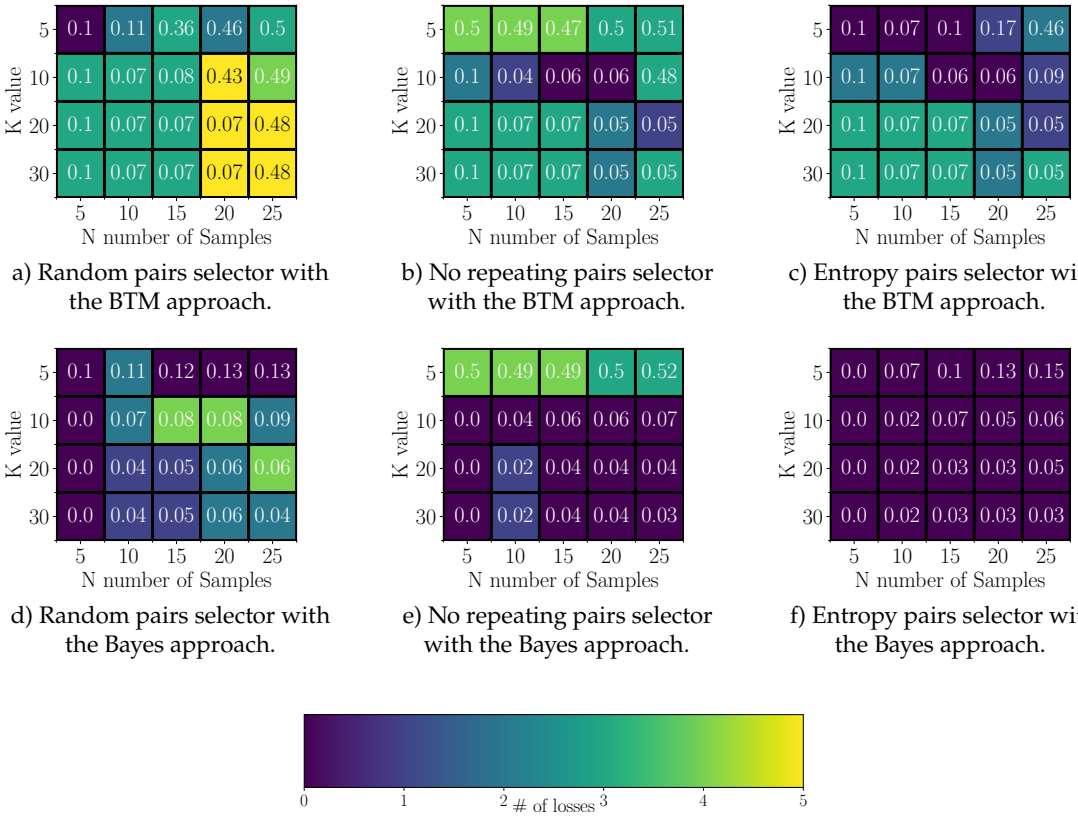
Figure 3.6: Convergence plots of the main current method for conducting CJ, a combination of the NR pairing method and BTM (Figure 3.6a)), and our novel entropy pairing method with BCJ (Figure 3.6b)). We can see that the BTM method, over time, hits an optimum level but then starts to deteriorate, while the entropy and Bayesian approach always gets more accurate with more data.

item in a pair would be the winner, which eventually misleads the BTM algorithm. In contrast, BCJ estimations consistently improved as more data became available.

In the following sections, we first discuss the performance of different methods in terms of τ distance. These results are shown in Figure 3.7. Here, we can see that overall, the Bayesian approaches performed better than BTM. However, the BTM with the entropy-picking method performed reasonably well compared to the other BTM combinations. It should be noted that to use the entropy-driven AL with BTM, we must construct Bayesian densities in the matrix $\mathcal{P}$.

In contrast, the Bayes and entropy picking method performed considerably better than the rest, with Figure 3.7f) showing that this combination was not beaten by any other combination method across all the experiments we conducted, demonstrating that it is significantly better or, worst case, performs the same as one of the other methods. Interestingly, this shows that our novel approach is better at generating a rank within a lower K value than suggested. Furthermore, the convergence plots in Figures 3.5 and 3.6 support this claim. Additionally, when the K value increases, it still performs well, which is irrelevant to the N value, as this does not affect its performance.

Therefore, overall, we can suggest that the Bayes version as a ranking method has done better, but the combination of Bayes and Entropy has done the best overall. Especially when comparing the current state-of-the-art approach (Figure 3.7b)) and our two novel approaches (Figure 3.7f)).



a) Random pairs selector with the BTM approach.

b) No repeating pairs selector with the BTM approach.

c) Entropy pairs selector with the BTM approach.

d) Random pairs selector with the Bayes approach.

e) No repeating pairs selector with the Bayes approach.

f) Entropy pairs selector with the Bayes approach.

Figure 3.7: An illustration of the statistical comparison of results of the random (3.7a), 3.7d)), no repeating pairs (3.7b), 3.7e)) and entropy (3.7c), 3.7f)) selection methods with BTM (*top row*) and Bayesian (*bottom row*) approaches for generating ranks. The plots show the number of times that a combination of a ranking method and a pair selection method has been the best, or equivalent to the best, with the darkest colour representing that it was not beaten by any other method for that configuration. The number in white shows the median performance over 50 repeats for the experimental configuration in the respective cell, with $BCJ^{\bar{E}}$ showing the best median performance in 18 out of the 20 distinct experiments.

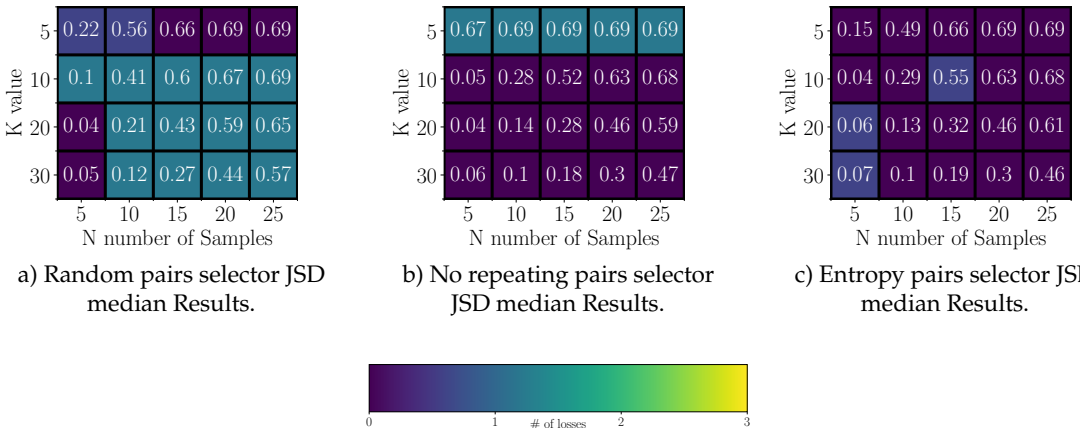
We note that in a real-world scenario, in the absence of information regarding target densities and expected ideal ranks, we cannot compute τ distances. In this case, we recommend using Figure 3.1 for investigating the current state of the preference PDF between any pair of items, and deriving the resulting rank distribution in Figure 3.2 (bottom row). One can also track the entropy reductions using Figure 3.4.

The results clearly demonstrate that the combination of BCJ and entropy-based pairing significantly outperforms traditional approaches, particularly at lower budget levels. From an educator's perspective, this means that fewer comparisons are required to achieve a high-confidence rank order, saving time and effort. This finding supports the overarching thesis objective of making robust, scalable assessment systems that remain grounded in practical classroom realities.

3.3.2 Efficacy in Rank Distribution Predictions

Due to the BCJ's ability to estimate the complete probability distribution over the rank of an item, we can compare the target densities from the items being compared. Again, in a real-world scenario, this comparison will not be possible, as we do not know the initial target distributions *a priori*.

Here, we use the JSD measure to identify the agreement between our BCJ estimate and the actual target distributions. For N items, we deduce N distributions over ranks and compare with their target counterparts. This comparison gives us N JSD values. We



a) Random pairs selector JSD median Results.

b) No repeating pairs selector JSD median Results.

c) Entropy pairs selector JSD median Results.

Figure 3.8: A comparison of the median JSD results over 50 repeats of 20 different experimental configurations for BCJ^R (*left*), BCJ^{NR} (*middle*) and BCJ^E (*right*).

take the worst JSD as reflective of the performance of the current rank distribution and track this throughout the BCJ process as a measure of progress.

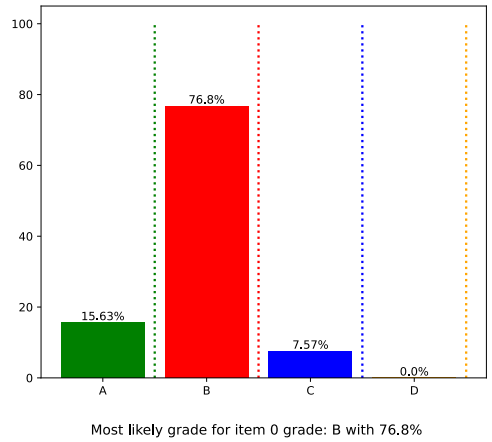
The results in Figure 3.8 show the efficacy of using different pair selection methods when used with BCJ. We see that for $K = 5$, using Entropy is the best strategy, with random being a close second. Essentially, when there is a lack of data with respect to the number of items being compared, random becomes competitive. However, it seems that no repeating pair strategy is the best for higher K values, with entropy beaten in three instances. Although it may be a good strategy with the synthetic targets we constructed, we would still recommend using the proposed uncertainty-based approach, i.e. entropy-driven AL, for larger Ns, as for unknown uncertainty densities over targets, no repeating pairs may not perform as well.

Unsurprisingly, comparing Figure 3.7 and 3.8, it is evident that BCJ is better at estimating the expected rank than the complete density of the rank distribution. For example, in Figure 3.7, with $N = 25$ and $K = 30$, BCJ^E has a median τ distance of 0.03, which means that only about 3% of all possible pairs, i.e. 9 out of 300, differ in order. In contrast, in Figure 3.8, the median of the worst matched item's rank density has a JSD of 0.46, which is far from the ideal match score of 0. It is reasonable to expect that with a larger budget on the number of paired comparisons, the rank agreement will improve.

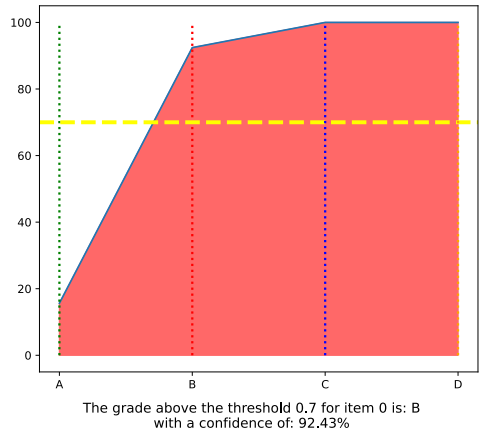
3.3.3 Assigning Grades

Different education systems grade assignments differently. For example, in England, exam boards use grades 9 to 1. In contrast, in the educational system in Wales, schools use the more traditional method of A^* to F, while vocational subjects in England and Wales use a Level 2 Distinction* to Level 1 Pass grading system. Typically, these grades are often assigned based on what *percentile* the work falls into compared to its peers, and these grades are ultimately what the assessors want to provide to the students. Therefore, it is important to be able to provide a possible grade based on the CJ results to help the assessors.

Typically, CJ scores are simply used and scaled to provide items with an absolute value between predefined upper and lower bounds. One possible approach to convert the rank information to a grade is to come up with a set of grade boundaries in terms of percentages of items that should get a certain grade. To the best of our knowledge, the only example of



a) Grade Probability distribution presented to the user.



b) The cumulative results of the probabilities and the threshold level set (shown in black dashed line) to be able to present the expected grade to the user.

Figure 3.9: A figure of the two methods used to present a predicted grade to the user. The panel on the *left* depicts the probability a student will get a particle grade, while 3.9b) the panel on the *right* shows the likely grade that meets the threshold level set by the user.

such an approach in practice uses national historical data to determine the grade boundaries in terms of percentages of items, as explained by Pinot de Moira *et al.* [149].

Taking inspiration from this, we propose using the probability densities over the rank of items to assign a grade to individual pieces of work. Given a discrete probability distribution over the rank of an item, we can compute the probability that an item's rank would be between two values as follows:

$$P(g \leq r_i \leq h) = \sum_{k=g}^{h} P(r_i = k), \tag{3.16}$$

where g and h are the boundary rank of the grade level. Using this, we can easily compute the probability that a piece of work lies between a range of ranks, and thus it can be interpreted with the notion of how many pieces of work should get the highest grades, and so on. This determination of grade is then entirely dependent on the assessor's decision on how many students should get what grade; for example, an assessor may decide that only the top 30% would receive a grade 9 (for an assignment submitted in England).

Figure 3.9 demonstrates this approach through an example of the outcomes after completing the CJ process. The teacher has decided that out of five pieces of work, one can receive a grade of A and B, two can receive a C, and one can receive a D. It gives us great

insight and therefore presents to the marker, for example, that item 3 (shown in the *left panel* of Figure 3.9) has a 15.63% probability of obtaining a grade A, 76.8% a B, 7.57% a C and 0% a grade D. Taking into account the cumulative probabilities, we can see that there is a $(15.63 + 76.8)\% = 92.43\%$ chance that this item would receive grade B or higher. If the assessor then decides on a *threshold of acceptability*, for example, 90%, to achieve a certain grade, we can assign grade B for this work. However, if the threshold was higher, e.g. 95%, the work would receive a grade of C, as then the cumulative probability would be at $(15.63 + 76.8 + 7.57)\% = 100\%$, which is higher than the threshold.

The ability to provide predicted grades is only possible due to our BCJ approach, which provides the probability distribution that an item will rank, as seen in Figure 3.2. We expect that such probabilistic reasoning renders the assessors greater control over the whole CJ process, with a high level of explainability.

Providing probabilistic grade distributions enables assessors to tailor grade assignments based on pedagogical intent and institutional policy, rather than relying solely on abstract rank scores. This respects teacher agency and supports flexible, criterion-referenced assessment design. It also adds a level of interpretability and fairness that current CJ systems often lack—enhancing the credibility and practical value of automated or semi-automated assessment tools in real educational settings.

3.4 Bayesian Comparative Judgement on a Real Comparative Judgement Dataset

To explore the real-world feasibility of the BCJ model in educational contexts, we now apply it to an established comparative judgement dataset. This allows us to evaluate how well our model performs outside synthetic conditions, especially in terms of uncertainty quantification and interpretability.

In 2018, Bramley *et al.* [156] used a round-robin approach for selecting pairs, much like the no repeating pairs approach used in this chapter, and demonstrated using GCSE English essay score data that adaptive CJ can incorrectly generate inflated confidence in their results. They performed three CJ assessments: study $1a$ was done using adaptive CJ; study $1b$ was done by using a pairing method of all-play-all (like round-robin for the same number of repeated evaluations), and a final one, denoted as study 2, was done using random pairings. For our demonstration with real data comparing BCJ and BTM

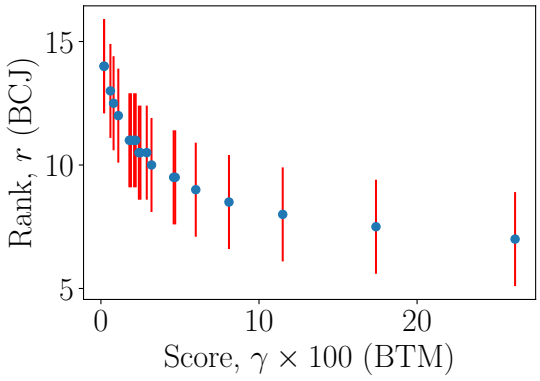


Figure 3.10: Comparison between the estimated ranks r_i using BCJ and scores $100 \times \gamma_i$ using BTM (see equation: 2.7). The blue dots show the expected rank $\mathbb{E}[r_i]$ versus the score, and the error bar (in red) shows the standard deviation of the predicted distribution over an item's rank. The full predictive distributions are shown in Figure 3.11. The higher the γ_i value, the better the item performed in the BTM ranking, and that corresponds to a lower expected rank, i.e. the better the item performed in the BCJ ranking, with a Kendall's τ rank correlation of over -0.97. The narrow difference between the expected ranks may indicate that the true performance difference between the items is likely to be low.

in this section, we selected the data from study $1b$. In this dataset, they used 18 judges with 20 distinct items of student work, which resulted in a total of 180 paired comparisons made by the judges, i.e. each judge assessed 10 pairs, and the SSR score was reported as 0.818, which is deemed as highly reliable. The scope and breadth of this dataset are similar to the synthetic experiments illustrated in earlier sections of this chapter, and, therefore, this dataset was selected for this demonstration.

The Kendall τ rank correlation coefficient between the BCJ rank vector $\mathbf{r}$ and BTM score vector $100\boldsymbol{\gamma}$ was -0.97319 with a p-value of 4.776×10^{-9} (which is practically zero), allowing us to reject the null hypotheses that the quantities are statistically independent. In other words, it shows that these two scores are almost perfectly anti-correlated in their estimations of ranks. In Figure 3.10, we clearly demonstrate this: the lower the predicted score from BTM, the higher the estimated rank from BCJ, as higher marks yield a lower rank, with 1 being at the top.

In Figure 3.10, we also show the standard deviation of the rank vector $\mathbf{r}$ with red vertical errorbars. For this experiment, the standard deviation for an item's estimated rank turns out to be $\sigma \approx 1.9$, and it is (practically) the same for every item's rank estimation. This is due to the distributions in Figure 3.11 being similar in shape and width. We

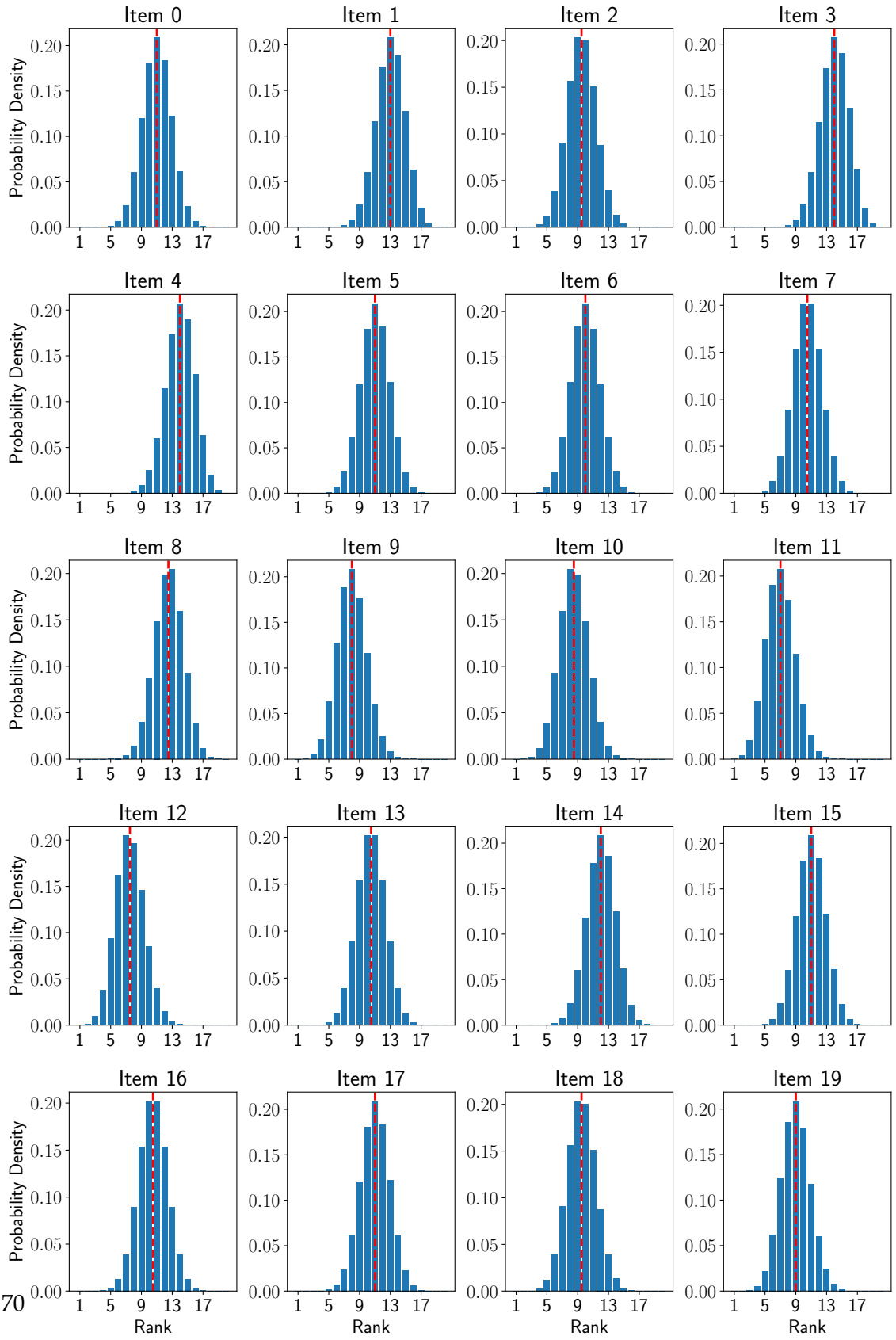


Figure 3.11: Illustration of the predictive probability distribution generated using BCJ along with the $\mathbb{E}[r_i]$ for each item i (depicted with red dotted vertical lines) using a real world dataset $1b$ from Bramley *et al.*. The experiment had an SSR score of 0.818, which is considered a respectable level for CJ as it is above the minimum of 0.7.

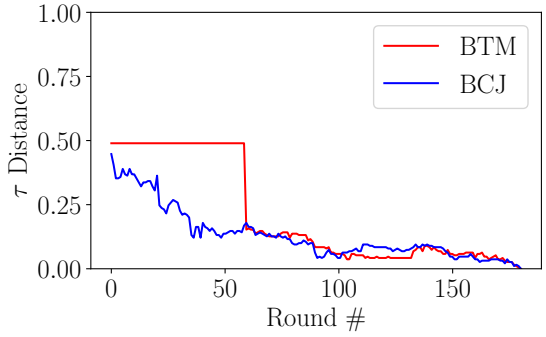


Figure 3.12: A comparison between the convergence of the BTM CJ (red line) and BCJ (blue line) against their respective final ranks. The BTM approach took ≈ 60 comparisons before generating a reasonable rank. Until this point, it produced a flat τ distance value of 0.5. Our Novel BCJ approach started to generate reasonable ranks even after the first comparison, and produced ranks with a τ distance in the region of ≈ 0.1 before the BTM τ distance started to improve.

attribute this to the fact that every pair had the same number of judgments. This also potentially indicates that individual assessors are fairly consistent in their judgements across items, and thus, the disagreements between assessors are consistent too; possibly as a consequence of the calibration exercise. However, to establish this, further experiments, both qualitative and quantitative, must be performed. In any case, there is enough signal in the data for us to identify differences between the expected ranks and derive an accurate rank order between the items, as confirmed by BTM's results, with the additional benefit of a clear depiction of uncertainties of predictions.

In Figure 3.12, we illustrate the convergence of the τ rank distance for BTM CJ, depicted by the red line, and BCJ, depicted by the blue line, against their respective final rank. We can see that the BCJ blue line starts to come down instantly after the first comparisons and continues to get closer to zero, while the BTM CJ approach stays at a τ score of 0.5 till ≈ 60 comparisons have been made, and then continues to drop. Before both end up reaching their final ranks, we can see that for the majority of the time, the BCJ blue line is below the red line. Therefore, showing a more consistent convergence in comparison to the BTM CJ approach.

It should be noted that the scores in BTM must be scaled to match any desired range, and then a grade can be derived based on pre-defined boundaries within that range. Whereas, in BCJ, we can provide predictions for ranks, which are immediately interpretable. Furthermore, we show how a grade can be assigned to an item based on relative, rather than absolute, performance, in Section 3.3.3.

The application of BCJ to real student data confirms its validity outside synthetic conditions, demonstrating both high agreement with traditional CJ results and enhanced interpretability through uncertainty estimates. The ability to show confidence intervals alongside rank positions adds a crucial layer of transparency. This supports the broader thesis claim that BCJ is not only a more technically robust model, but one that aligns better with educators' needs for trust, fairness, and explainability in high-stakes assessment contexts.

3.5 Measuring Reliability

Although the BCJ method is attractive for its speed and accuracy, a key open question is how to assess its reliability. In this chapter, we address this by examining inter- and intra-rater agreement as indicators of reliability.

Previously, we suggested that one can track the maximum entropy across all the possible pairs due to the Beta posterior distribution in each, and when it is *sufficiently* low, one can stop selecting further pairs [223]. However, an entropy value can be difficult to interpret, and only makes sense as a relative measure, making it challenging to measure and communicate reliability, or to devise a stopping criterion for pair selection.

One feature of estimating the posterior Beta distribution over the preference between two items is that it directly encapsulates the level of agreement between the decisions that were made about a particular pair. This means when a pair truly divides the crowd (be it inter or intra rater), the probability Beta posterior distribution would have an expected value of 0.5, where 0 represents perfect agreement on an item losing and 1 represents the same item winning; see Figure 3.13 for an illustration of these possible cases.

With this, we can formulate measures of reliability that diverge from the expected highest level of disagreement of 0.5. Given that the most likely value of a Beta posterior is the mode, we can, firstly, define it to capture the divergence of the mode from 0.5. Noting that the direction of divergence does not matter, we define the mode agreement percentage (MAP) as follows:

$$MAP(\alpha_{post}, \beta_{post}) = \frac{|m(\alpha_{post}, \beta_{post}) - 0.5|}{0.5} \times 100\%, \tag{3.17}$$

where the mode $m(\alpha_{post}, \beta_{post}) = \frac{\alpha_{post}-1}{\alpha_{post}+\beta_{post}-2}$ with α_{post} and β_{post} are the posterior parameters for the Beta density over preference for a pair.

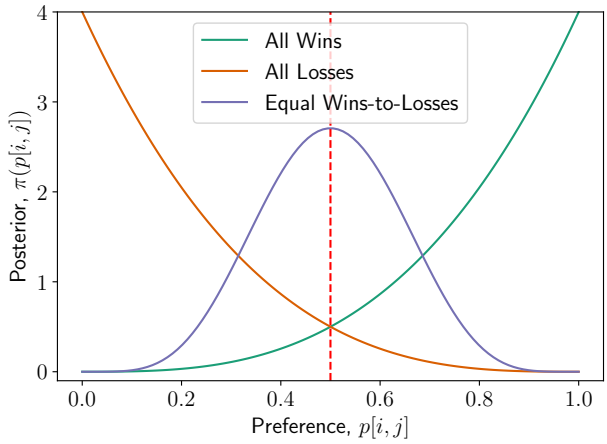


Figure 3.13: An illustration of the posteriors under different levels of agreements. When all ratings agree, on either all wins (shown in green) or all losses (shown in orange), for item i compared to item j, the densities skew towards 1 or 0, respectively, with the corresponding most likely predicted outcome being close to 1 or 0. On the other hand, if we have an equal number of wins and losses, i.e. the highest level of disagreements between ratings, we get the purple density with the most likely outcome being 0.5 (depicted with the red dashed vertical line). Here, we assumed 4 comparisons have been made; with more comparisons, variance would reduce given the assumptions for outcomes.

While this provides an intuitive avenue to measure reliability, it does not appropriately incorporate the uncertainty from the paucity of comparison data per possible pairing. To capture the uncertainty in a measure, we, therefore, propose to calculate the expected agreement percentage (EAP) as follows:

$$
\begin{aligned}
EAP(\alpha_{post}, \beta_{post}) &= \kappa \int_0^1 p^{\theta_1} (1-p)^{\theta_2} |p - 0.5| \, dp \\
&= -\kappa \left[\frac{0.5\Gamma(\theta_1+1)\, {}_2F_1\left(\begin{matrix} -\theta_2, \theta_1+1 \\ \theta_1+2 \end{matrix} \middle| 1 \right)}{\Gamma(\theta_1+2)} - \frac{1.0\Gamma(\theta_1+2)\, {}_2F_1\left(\begin{matrix} -\theta_2, \theta_1+2 \\ \theta_1+3 \end{matrix} \middle| 1 \right)}{\Gamma(\theta_1+3)} \right] \\
&+ 2\kappa \left(\frac{0.25 0.5^{\theta_1}\Gamma(\theta_1+1)\, {}_2F_1\left(\begin{matrix} -\theta_2, \theta_1+1 \\ \theta_1+2 \end{matrix} \middle| 0.5 \right)}{\Gamma(\theta_1+2)} - \frac{0.25 0.5^{\theta_1}\Gamma(\theta_1+2)\, {}_2F_1\left(\begin{matrix} -\theta_2, \theta_1+2 \\ \theta_1+3 \end{matrix} \middle| 0.5 \right)}{\Gamma(\theta_1+3)} \right),
\end{aligned}
\tag{3.18}
$$

where, $\kappa = \frac{\Gamma(\alpha_{post}+\beta_{post})}{0.5\,\Gamma(\alpha_{post})\Gamma(\beta_{post})} \times 100$, $\theta_1 = \alpha_{post} - 1$, and $\theta_2 = \beta_{post} - 1$, with $\Gamma(\cdot)$ is the Gamma function and ${}_2F_1(\cdot)$ is the Gaussian hypergeometric function. We validated this result through simulation.

These formulations for MAP and EAP around 0.5 relate to percentiles over preferences. Specifically, the MAP (or EAP) metrics indicate how far the metric value is from the middle, on both sides, and thus inform us of the range beyond which we currently have the metric. We can calculate the lower bound of the range with $l = 0.5 - \frac{0.5\ \text{MAP}}{100}$ and the upper bound of the range with $u = 0.5 - \frac{0.5\ \text{MAP}}{100}$. For instance, a 50% MAP means that the mode resides outside the range between $l = 0.25$ and $u = 0.75$. In terms of EAP, since this is integrated over the uncertainty in the density, a 50% EAP would mean that there is enough volume to push the expected value of the agreement percentage beyond the range between $l = 0.25$ and $u = 0.75$. Hence, we can devise a stopping criterion based on the desired level of confidence, and thus enforce a range for this "null space".

Alternatively, the assignment owner can decide the lower and upper bounds of this "null space" and then compute the threshold required for the minimum MAP or EAP before stopping further data collection. For example, if they wanted the width of the "null space" to be 95%, they could define a range between $l = 2.5\%$ and $u = 97.5\%$, which would be equivalent to a threshold of 95% on MAP or EAP (whichever they were tracking for this purpose).

In terms of the choice of MAP or EAP, we noted that they both are useful in different ways. MAP provides an intuitive indication of where the mode is, but because it does not consider the level of existing uncertainty, it can be overly optimistic. On the other hand, EAP provides a more comprehensive metric of reliability that incorporates the amount of information at hand as we integrate over the uncertainty in the density. For example, consider a case when an item always wins in a pair. The mode would quickly shift towards the right even with a few wins, the mode would quickly shift towards the right, like the green line depicts in Figure 3.13. However, the variance does not diminish so rapidly. Hence, the MAP will show a rather instantaneous shift towards 100%, but EAP would only do so when there are numerous comparisons and all indicate wins for the item: see Figure 3.14. When the observations fluctuate between wins and losses, these instances shift in mode, making MAP fluctuate more acutely, especially when there is limited data; see Figure 3.15.

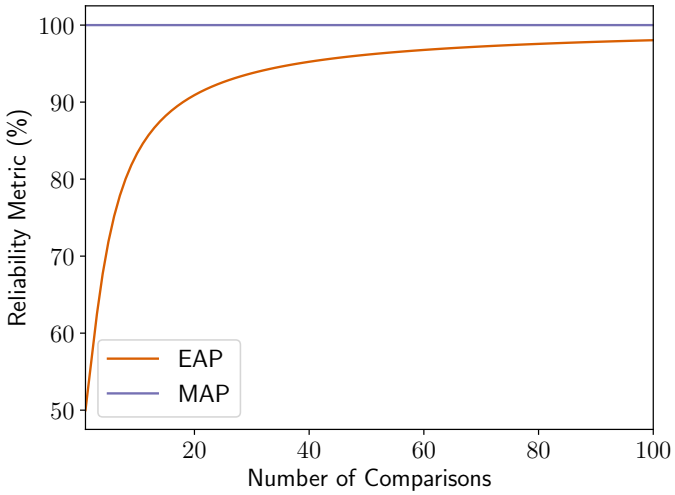


Figure 3.14: An illustration of EAP increasing slowly (shown in orange) as we observe an item winning at every comparison with another specific item to reflect the decreasing uncertainty over comparisons. Whereas MAP, shown in purple, is overoptimistic and quickly gets to near 100%, even with a few observed wins.

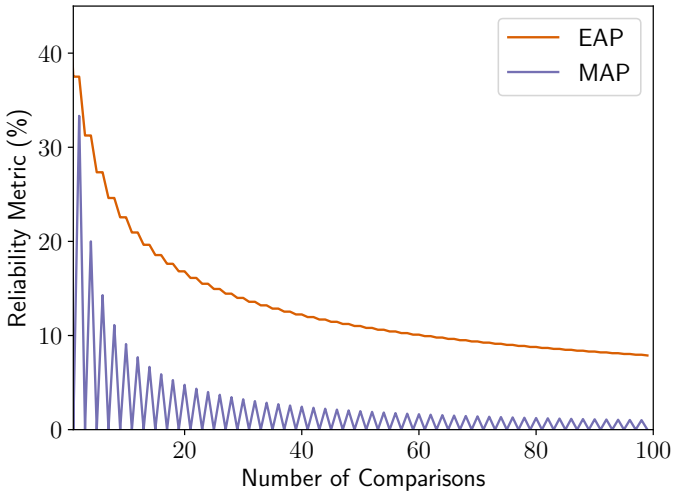


Figure 3.15: An example of EAP being more stable when there is conflicting information, with an item only winning every second comparison against a particular item. MAP fluctuates rapidly, but with sufficient data, the overshoots are small (depicted in purple).

It should be noted that the decision to prefer one over the other in paired comparison may be made by the same individual at different times or different individuals (either synchronously or asynchronously), and the Bayesian machinery here would treat them the same way. Thus, both of these reliability metrics account for both inter- and intra-rater reliabilities, depending on the context of data collection.

3.5.1 Assessing Reliability and Integrating Principal Marker Interventions

In this section, we begin our investigation into assessing reliability using single-criterion BCJ. For an arbitrary instance of the DREsS dataset with $N = 10$ and a budget multiplier $K = 10$, we run BCJ for $N \times K = 100$ simulated pairwise comparisons, driven by entropy-based selection. We record the final τ score with respect to the ground truth, along with the MAP and EAP scores for each pairwise comparison.

Unlike a single SSR score, the proposed MAP and EAP metrics offer a more nuanced perspective on assessor agreement. These metrics enable the identification of specific item pairs that contribute to disagreement, providing actionable insights that a single, aggregate reliability score cannot capture. By using MAP and EAP, we gain a clearer understanding of uncertainty at the pairwise level (see the upper triangle of Figure 3.16 for a visual illustration). Importantly, this analysis pertains to reliability – whereas ranking accuracy is determined solely by the Bayesian estimation of rank distributions, which is theoretically optimal given the available data.

To showcase the practical value of the EAP metric, we simulated an intervention by a principal moderator (PM). By flagging item pairs with low agreement (EAP < 50%), the PM could concentrate on the most contentious comparisons and make a judgement on which item should be considered superior. The chosen item in each pair was then assigned an artificial win count of 1000, indicating strong confidence in the decision and effectively removing it from future selection via entropy-based methods.

Figure 3.16 illustrates this process: the upper triangle highlights low-agreement pairs (EAP < 50%) with green boxes, while the lower triangle displays the updated EAP scores and distributions following the PM's intervention.

The ground-truth target ranking for the items was:

$$\langle 7, 0, 6, 2, 5, 4, 3, 8, 9, 1 \rangle,$$

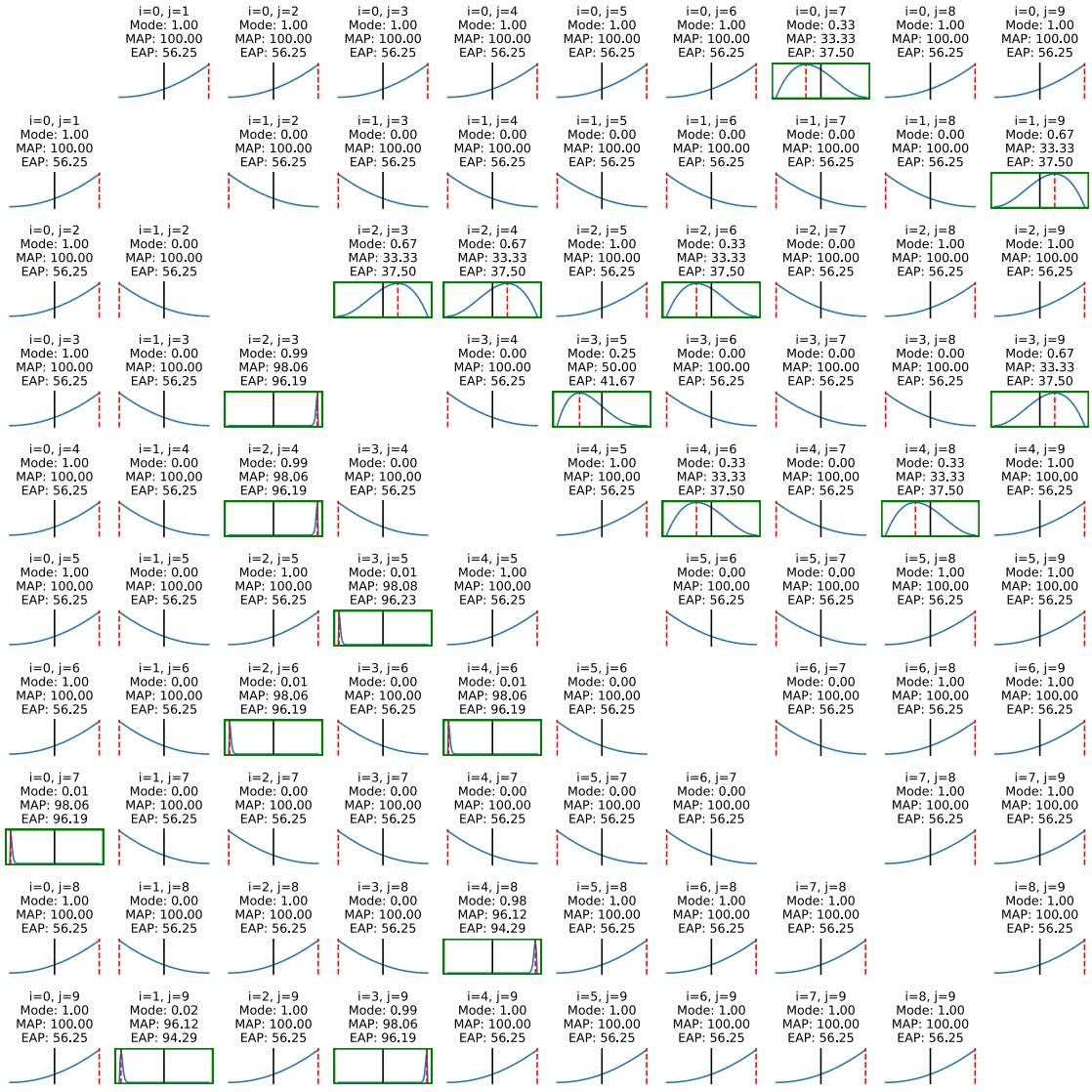


Figure 3.16: MAP, and EAP scores for each pairwise comparison in the DREsS dataset ($N = 10$, $K = 10$). Comparisons with EAP scores below 50% were flagged and reviewed by a PM to identify items causing disagreement (shown within green boxes). The PM then selected the winner, and we biased the respective preference distributions accordingly. The upper triangle displays the original decisions prior to intervention, while the lower triangle reflects the updated outcomes after moderation.

while the estimated ranking prior to intervention was:

$$\langle 7, 0, 6, 2, 4, 5, 8, 3, 1, 9 \rangle.$$

This reveals three misordered pairs: $\langle 5, 4\rangle$, $\langle 3, 8\rangle$, and $\langle 9, 1\rangle$. Following the intervention, two of these — $\langle 5, 4\rangle$ and $\langle 9, 1\rangle$ — were corrected. The pair $\langle 8, 3\rangle$ remained unresolved, as item 3 had only been compared with items 2, 5, and 9, and was consistently judged superior. Item 8 was also deemed better than those, leaving no new information on the direct comparison between items 3 and 8 to inform their relative ranking. Nevertheless, this targeted intervention improved ranking accuracy, reducing the τ score from 0.07 to 0.04.

Importantly, we do not account for probabilistic transitivity due to the independence assumption in BCJ, which aligns with findings in human decision-making behaviour [180]. While incorporating inter-item dependencies could enhance accuracy, it would significantly complicate the probabilistic modelling.

3.6 Conclusions

The results from both synthetic and real-world experiments indicate that BCJ is a viable, scalable, and educator-focused alternative to traditional CJ methods. This chapter demonstrates how Bayesian modelling not only improves technical performance but also supports the overarching thesis aim: to develop more transparent, efficient, and trustworthy approaches to student assessment.

Marking and assessing the work of students is an important element of education. However, it takes a long time and can be inconsistent, especially because we are not great at assessing absolute quality. Furthermore, we are beginning to see the use of generative AI tools in education and their potential impact on various forms of assessment and associated practices [114, 224].

With the introduction of CJ, this has helped alleviate a lot of the quality issues in principle, but it does come with its own issues. One of the issues is that the paired comparison rank order starts to deteriorate, making the whole model's fit somewhat collapse. Also, it is not easy to determine how many comparisons are enough. As the study has shown that the τ distance score gets worse as the value of K gets larger. While the recommended minimum number of comparisons is $N \times 10$, this study has shown that it struggles after $N \geq 20$, showing that a larger K is required as at the suggested minimum the current CJ with BTM struggles to rank accurately, with results showing that when $N = 20$ a K value of 20 is required to start getting close to the desired rank.

Nonetheless, our novel BCJ approach does not suffer from this issue, as the more comparisons we make, the more accurate it gets.

Most importantly, there are issues around using any current form of CJ as a replacement for marking, as the outcome is less transparent [147]. During the design of our new BCJ approach, we focused on addressing the issue of transparency by being able to provide information to the user about how the algorithm has come up with its rank decisions, as well as allowing the user to give input into how it generates the grades, as well as giving the information on how it predicted what it has predicted. Therefore, rendering greater transparency compared to the standard approach, and it is computationally affordable too.

In a future chapter, we intend to use the proposed approach to the CJ process with educators. In both quantitative and qualitative manners, we will seek to answer what works and what doesn't and how BCJ is received in a real-world study.

In summary, BCJ addresses key limitations of traditional CJ by introducing a probabilistic, transparent, and efficient ranking model that has high accuracy and maintains stability as comparisons increase. However, while BCJ effectively provides holistic rankings, it lacks the criterion-specific granularity required for detailed, actionable feedback in a rubric-based assessment.

To address this limitation, the next chapter introduces Multi-Criterion BCJ (MBCJ), an extension of BCJ that enables detailed evaluation across multiple learning outcomes. This approach bridges the gap between holistic ranking and structured feedback, further enhancing the transparency and practical utility of CJ in educational contexts.

Chapter 4

Multi-Criteria Bayesian Comparative Judgement

Building on the successful application of BCJ for holistic assessment, this chapter explores its extension to multi-criteria assessment. Educators often require insights into specific learning objectives (LOs) to provide detailed, actionable insights to assessors and students. The proposed Multi-Criteria Bayesian Comparative Judgement (MBCJ) framework addresses this need by enabling simultaneous holistic and criterion-specific assessment, using entropy-based active learning to enhance efficiency while maintaining transparency in the assessment process.

4.1 Introduction

Humans are generally better at relative judgments than absolute scoring, which underpins the logic of CJ introduced by Thurstone [135, 225]. CJ infers rankings from pairwise comparisons, often modelled using the Bradley–Terry model (BTM) [226]. While frequentist approaches have traditionally dominated the Bradley–Terry literature hunter2004mm, Bayesian alternatives have been increasingly explored in recent work caron2012efficient, offering improved uncertainty handling. However, interpretability remains a challenge since only comparisons—not true scores—are observed.

Active learning has been proposed to optimise pair selection [178, 179], but most methods remain heuristic or frequentist. This chapter introduces Bayesian Comparative Judgement (BCJ) [223], which models comparisons as Bernoulli trials, avoiding restrictive

assumptions and naturally handling stochastic, imperfect judgements. BCJ uses Beta distributions to quantify uncertainty and supports entropy-based active learning for efficient pair selection across multiple criteria, outperforming BTM-based methods empirically.

Beyond validity, reliability is critical, given inherent assessor variability [180, 227]. Traditional metrics like SSR [228] have limitations, and no widely accepted Bayesian equivalent exists yet. Finally, while CJ decisions are holistic, multiple latent attributes often influence judgements [229], a complexity addressed in multidimensional extensions since Hefner's early work [230]. In many real-world judgement scenarios, however, the relevant criteria and their contributions are explicitly defined in advance. For instance, in educational assessment, students are evaluated using a predefined rubric that specifies distinct criteria, each contributing to the final mark with predetermined weights. In such cases, it would be natural to collect pairwise comparisons at the level of individual criteria. This has led to approaches where CJ is used for multiple focuses, allowing assessors to create independent rankings for each criterion (e.g., separating focus on structure from focus on argumentation) [231, 232]. This approach has been referred to as CJ dimensions in some literature [232]. Following this multi-focus ranking, another study has conducted a correlational experiment to determine which criteria rankings are significantly correlated [231]. However, in order to gain these insights, the CJ process was executed as many times as required for each criterion being examined [231, 232]. Yet, to the best of our knowledge, an approach that looks to combine all criterion rankings into one holistic ranking has not been explored within the traditional CJ framework, which has historically emphasised holistic comparisons of the work as a whole.

Although some existing methods attempt to infer latent dimensions from holistic choices, it may be difficult to align these inferred dimensions with specific, predefined criteria. Moreover, they lack a principled mechanism for aggregating criterion-level decisions into a single, coherent ranking.

To address these limitations, this Chapter builds on BCJ and makes the following key contributions:

- We propose new methods for estimating overall ranks and associated predictive uncertainties from pairwise comparisons made per criterion. This framework, which we term Multi-Criteria BCJ (MBCJ), enables criterion-specific ranking and uncertainty estimation.

- We show how a holistic entropy can be calculated to drive the selection of the most informative pair to be evaluated next in MBCJ.

- We demonstrate, for the first time, that MBCJ performs comparably to standard BCJ in experiments using real assessment data, while providing finer-grained insights into item preferences across individual criteria.

This chapter builds on the foundations of CJ and BCJ introduced earlier by addressing one of their key limitations: the lack of detail regarding performance across individual learning outcomes. By introducing a multi-dimensional extension to BCJ, we aim to retain the benefits of holistic comparison while enabling more transparent, criterion-specific insight—an essential feature for practical classroom assessment and feedback.

4.2 Multi-Criteria Bayesian Comparative Judgement

Suppose an assignment is evaluated against D learning outcomes (LOs). The assignment owner specifies a weight vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_D)^\top$, where each $\lambda_d \in [0, 1] \subset \mathbb{R}$ indicates the contribution of the dth LO to the overall mark, with the constraint that $\sum_{d=1}^{D} \lambda_d = 1$. Given a score $\gamma_{i,d}$ for item i on LO d, the overall score for item i is computed as:

$$\gamma_i = \sum_{d=1}^{D} \lambda_d \gamma_{i,d}. \tag{4.1}$$

While this weighted aggregation is standard in rubric-based assessment, conventional CJ methods typically rely on holistic comparisons. This can obscure valuable insights into how an item performs across individual LOs. Although CJ provides fast and accurate overall rankings, assessors may find it difficult to revisit specific items and offer targeted feedback to learners on individual criteria.

To adapt CJ for multi-criteria rubrics, we propose a framework in which decisions are made independently for each LO. This requires suitable methods for combining these decisions into a unified overall ranking. Additionally, we introduce a strategy for selecting the most informative pairwise comparison by considering uncertainty across all LO-specific evaluations. In the following section, we present novel methods to address these challenges.

4.2.1 Extension to Rank Generation

To generate a combined final rank for each item, we propose two approaches: one that aggregates LO-specific ranks, and another that merges LO-specific preference distributions, represented by Beta posterior densities.

4.2.1.1 Mixture of Component Ranks

For dth learning outcome LO_d, the rank distribution for item i is denoted by $P(r_{i,d} = a)$, where $a \in [1, N] \subset \mathbb{N}$. Given a predefined weight vector $\mathbf{l}$, we can construct a mixture model to combine these LO-specific rank distributions as follows [233]:

$$P(r_i = a) = \sum_d \lambda_d P(r_{i,d} = a). \tag{4.2}$$

The corresponding expected rank for item i is then given by:

$$\mathbb{E}[r_i] = \sum_d \lambda_d \mathbb{E}[r_{i,d}]. \tag{4.3}$$

Figure 4.1 provides a visual illustration of the expected ranks $\mathbb{E}[r_{i,d}]$ for each LO. By combining these values using a weighted sum, we obtain the overall expected rank for item i.

A central challenge in this approach lies in deriving the overall preference distribution $\pi(i \succ j)$ between items i and j that corresponds to these combined rank distributions. Without a likelihood-based model — such as the one employed in BTM — this estimation becomes non-trivial.

4.2.1.2 Mixture of Component Preferences

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 [233]:

$$\mathcal{F}(i \succ j) = \sum_d \lambda_d \mathcal{F}_d(i \succ j), \tag{4.4}$$

where λ_d denotes the contribution of LO_d to the overall mark. We provide an illustration in Figure 4.2 of such a mixture CDF.

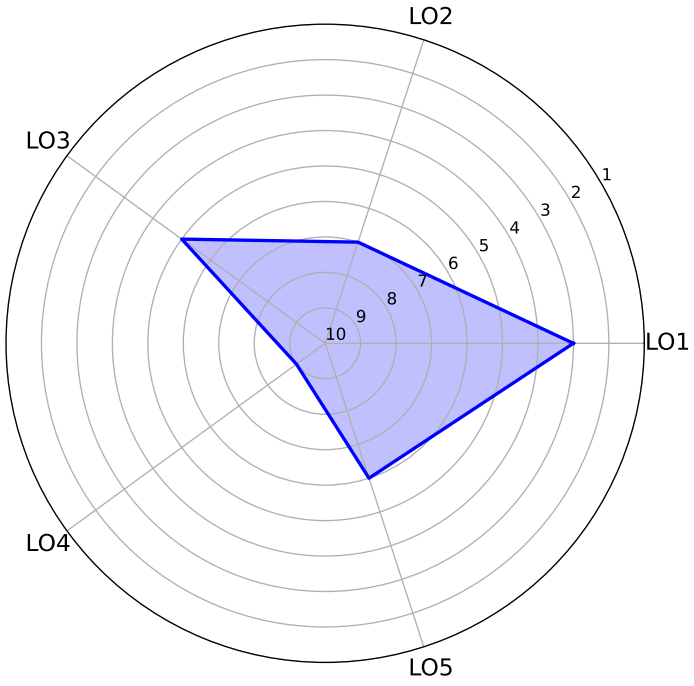


Figure 4.1: A radar plot depicting the ith item's $\mathbb{E}[r_{i,d}]$ performance across five LOs, enabling more transparency and detail on where this item performed well and where it did not. Therefore, it enables educators to identify areas where this candidate may need personalised intervention. Furthermore, it provides more insight than a traditional CJ rank would offer to the educator.

This formulation allows direct computation of the overall preference probability that item i dominates item j, using Equation (4.4). It also enables derivation of the full probability distribution over ranks for each item. Methodologically, this approach is advantageous, as both preference and rank distributions can be efficiently obtained using existing mechanisms.

However, exact calculation of overall preference and rank distributions becomes susceptible to combinatorial explosion when ranking a large number of items. To address this, a Monte Carlo (MC) sampling approach may be beneficial. For mixture of preference distributions, the mth sample p_m can be drawn as follows:

$$p_m \sim \pi_q(i \succ j) \mid z_m \sim U(0,1) \wedge z_m \in \left[\sum_{d=0}^{q-1} \lambda_d, \sum_{d=0}^{q} \lambda_d\right], \tag{4.5}$$

where z_m is the mth random number sampled from the uniform distribution $U(0,1)$, $\pi_q(\cdot)$ is selected based on the value of z_m, and $l_0 = 0$. A win for item i can be simulated

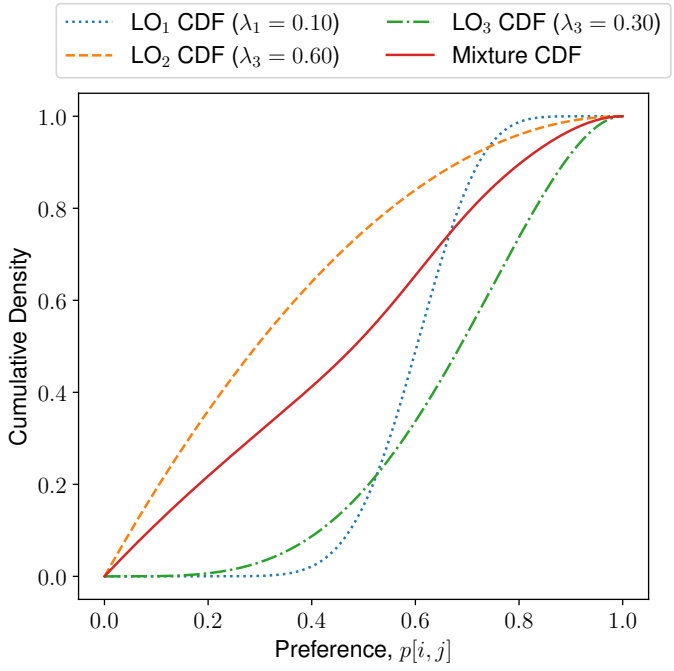


Figure 4.2: A visual illustration of how LO-specific preference distributions are combined using a weighted sum of their CDFs. In this example, three LOs are shown in blue, orange, and green, corresponding to the weight vector $\boldsymbol{\lambda} = (0.1, 0.6, 0.3)^\top$. The resulting mixture CDF, shown in red, is not a standard Beta distribution but effectively reflects the contributions of the individual components.

by rounding the sampled proportion to the nearest integer, $x_m = \lfloor p_m \rceil$. The total number of wins and losses across all pairwise comparisons, aggregated over all preference distributions, can then be used to estimate rank distributions and compute the expected overall rank for each item.

4.2.2 Extension to Pair Selection

Differential entropy generalises the classical concept of discrete entropy to continuous random variables, offering a measure of uncertainty associated with a probability distribution [234]. For the Beta distribution – defined on a finite interval $[0, 1]$ and parameterised by two shape parameters – differential entropy reflects a measure of uncertainty over that interval, as discussed in Section 2.2.3.3.

In the multi-dimensional case, each preference distribution $\pi_d(i \succ j)$ between items i and j for LO_d is modelled as a Beta distribution $\mathcal{B}(\alpha_{\text{post}}, \beta_{\text{post}})$. Assuming independence across LOs, the total entropy across all D LOs is then computed as [235]:

$$H(\pi(i \succ j)) = \sum_d H(\pi_d(i \succ j)). \tag{4.6}$$

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

$$(i, j) \leftarrow \arg \max_{(i,j) \wedge i \neq j} H(\pi(i \succ j)). \tag{4.7}$$

4.3 Experimental Setup

In this chapter, we aim to experimentally investigate the following questions:

(Q1) Which combinations of ranking and selection strategies perform best in multi-criteria CJ when weights are fixed?
(Results discussed in Section 4.4.1)

(Q2) Which approaches remain robust under varying weight configurations?
(Results discussed in Section 4.4.2)

This section begins by outlining the strategy variants included in our comparison, followed by an overview of the real-world datasets used. We then describe the simulation process for modelling assessor decision-making and conclude with a summary of the performance evaluation methodology.

4.3.1 Strategies Under Scrutiny

We explore six strategy combinations formed by pairing two ranking methods, Ranking Mixture Model (Section 4.2.1.1) and Preference Mixture Model (Section 4.2.1.2), with three pair selection techniques: entropy-based (Section 4.2.2), random, and NRP. These combinations are evaluated against the baseline BCJ ranking method with entropy-driven selection, as proposed in chapter 3.

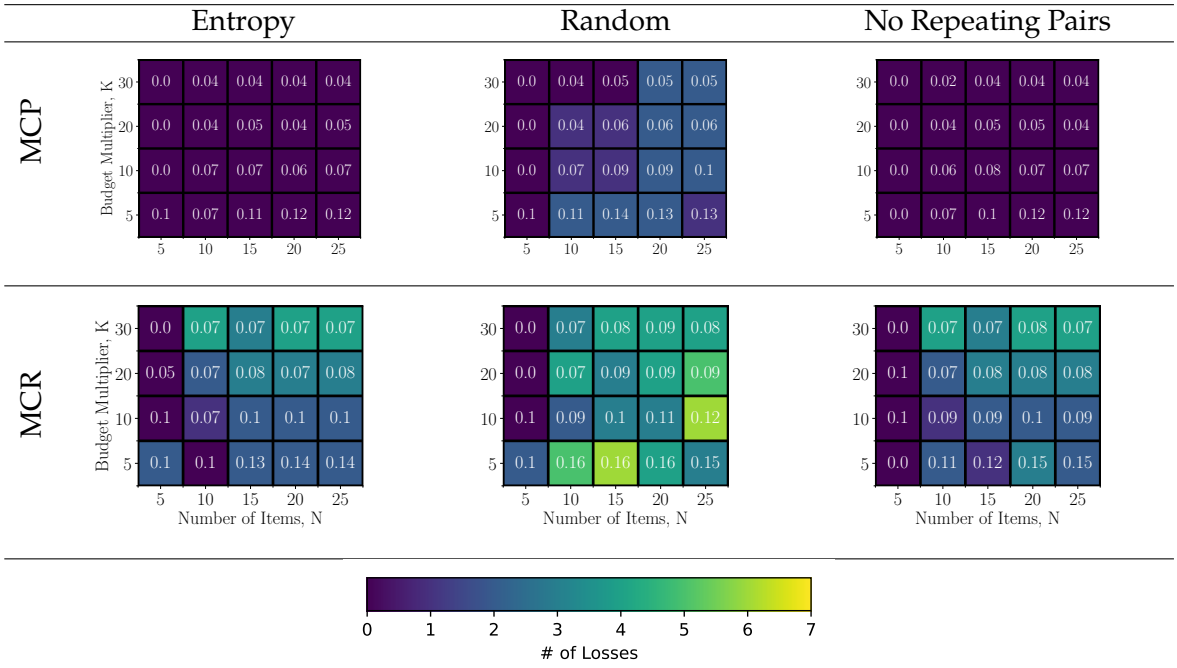


Figure 4.3: Statistical comparison of results from the Wilcoxon rank-sum test on the DREsS dataset for multi-criteria strategies based on the mixture of component ranks (MCR) and the mixture of component preferences (MCP). Each strategy in a panel is identified by the row and column labels. Each cell is coloured according to the number of items (horizontal axis) and the budget multiplier K (vertical axis). The colour of each cell reflects how often a particular strategy was outperformed by another competing strategy: darker colour indicate stronger performance (fewer losses), while lighter colour indicate weaker performance (more losses). The number shown in white within each cell represents the respective median performance. An MCP based strategy incorporating the NRP pair selection method demonstrates the best overall performance across the experiments with this dataset, with the entropy method a close second.

4.4 Results and Discussion

4.4.1 Identifying the Best Strategy

For every strategy – defined as a combination of ranking and pair selection methods – we conduct 50 repeated runs for a given budget $N \times K$, recording the final τ scores with respect to the ground truth. These results allow us to compare strategies using the Wilcoxon rank-sum test and identify the statistically superior approaches.

In Figure 4.3, we illustrate the results for the DREsS dataset. Among the multi-criteria strategies, MCP with entropy and NRP consistently outperform the others, as they are never beaten by any competing strategy. It should be noted that MCR, regardless of

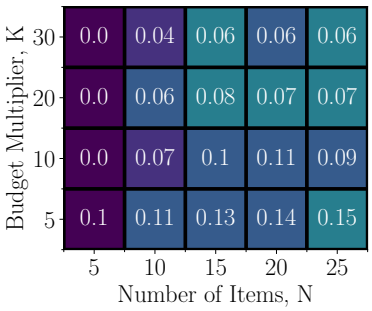


Figure 4.4: An illustration of the statistical comparison results for single-criterion (holistic BCJ with entropy-based pair selection) versus multi-criteria strategies. The colour of each cell represents how many times a given strategy was outperformed by others. Each cell displays the corresponding median performance in white. For $N = 5$, the BCJ strategy performs comparably to the best-performing strategy. However, for $N \geq 10$, there is at least one other strategy that consistently outperforms BCJ.

the pair selection method, performs reasonably well beyond $N = 5$, with the weakest performance observed for MCR using random pair selection. For the BCJ with entropy approach, shown in Figure 4.4, we observe that it is outperformed by at least one other strategy when $N \geq 10$.

For the BU dataset, which features tighter marking tolerance, we again observe that MCP is the superior approach among the multi-criteria strategies. In this case, the entropy-based pair selection method performs best, being beaten only once at $N = 5$ and $K = 30$ (see Figure 4.5). In this instance, the single-criterion BCJ strategy also performs well (see Figure 4.6), and is only outperformed for $N \in \{10, 20\}$ with $K = 5$, presumably due to limited preference data available at lower budget levels.

Overall, we observe that the strategy combining MCP ranking with entropy-based pair selection performs best across both datasets. While the NRP method paired with MCP also performs well on the DREsS dataset, we attribute this to the greater pair-wise uncertainty due to the tolerance level associated with that dataset. Typically, entropy selects the most informative pair. When the preference differences between items are clear (i.e. more certain), those pairs are unlikely to be revisited frequently. This creates a distinction between entropy-based pair selection, which targets the most informative comparisons, and the NRP method, which iteratively revisits all pairs in a round-robin fashion. However, in cases of high uncertainty, both entropy and NRP tend to behave similarly, as more frequent revisiting of uncertain pairs becomes necessary.

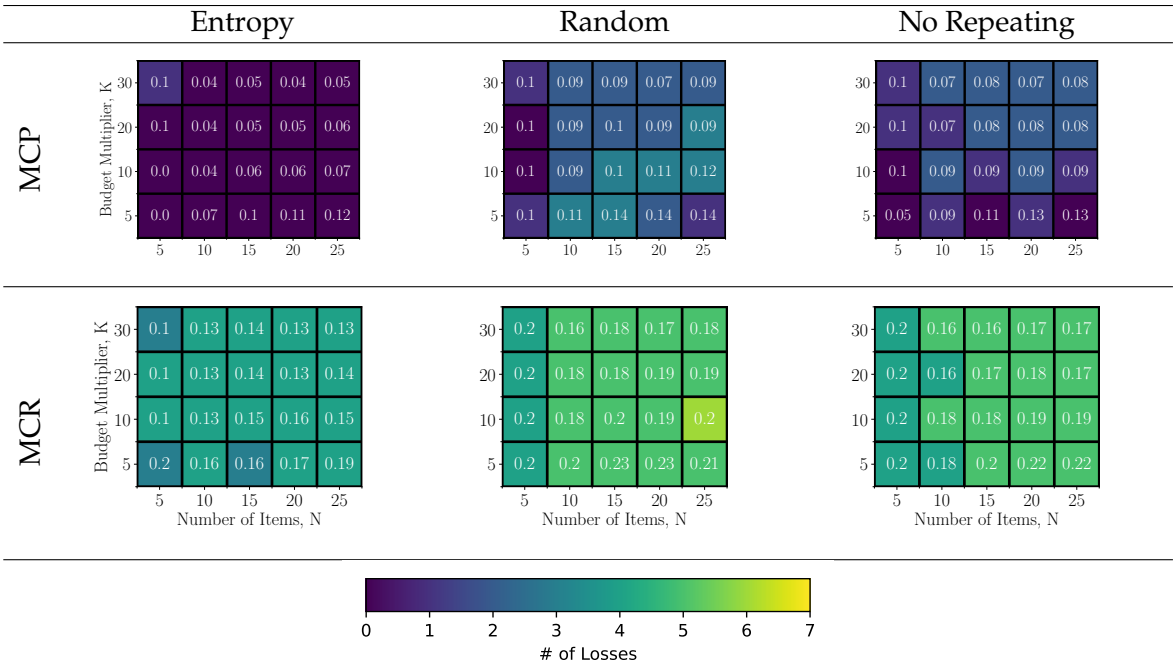


Figure 4.5: An illustration of the statistical comparison results from the Wilcoxon rank-sum test for multi-criteria strategies based on the mixture of component ranks (MCR) and the mixture of component preferences (MCP) on the BU dataset. The plots show the number of times each combination of ranking method and pair selection method was the best, or equivalent to the best. The darkest colour indicates that the strategy was not beaten by any other method for that configuration, including comparisons against the single-criteria BCJ strategy using the standard entropy-based pair selection method. The white number in each cell indicates the median performance for that category. The MCP strategy demonstrates the strongest overall performance across the experiments, being beaten only once at the $N = 5$, $K = 30$ configuration.

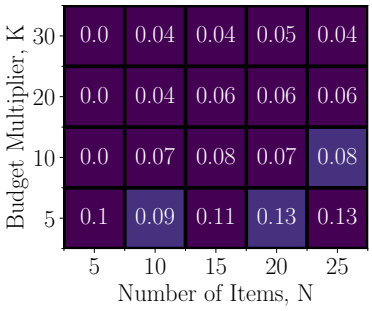


Figure 4.6: An illustration of the statistical comparison of results of the Wilcoxon rank-sum test for the level 4 undergrad dataset of the BCJ and entropy picking methods against the multi-criteria BCJ and other picking methods. We can see for this dataset apart from $N = 10$ and $N = 20$ for the K value 5, this approach was not dominated by any of the other combinations.

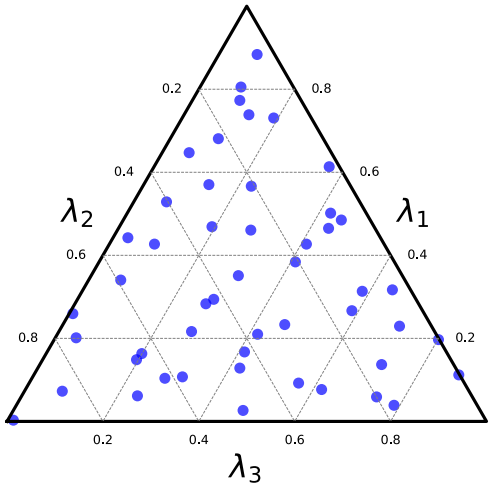


Figure 4.7: An illustration of the QMC weights transformed onto a simplex which is used for testing the robustness of the approaches.

In terms of the single-criteria BCJ strategy, it generally performs well, particularly when uncertainty in overall preferences is low. However, by design, it lacks the richness provided by LO-specific preference information, which may limit its effectiveness in real-world assessment scenarios; especially from a feedback perspective.

4.4.2 Robustness to Varying Weight Configurations

Both datasets come with predefined weights (as discussed in Section 2.4.1). Accordingly, we define a random performance vector $\tau(s \mid \boldsymbol{\lambda})$, where s represents a strategy and $\tau(\cdot|\boldsymbol{\lambda})$ is the performance metric for a specified weight vector $\boldsymbol{\lambda}$. While we observe clear benefits from the MCP strategy with entropy-based pair selection, a natural question arises: what happens if we vary the weights? Would the conclusions remain the same?

Addressing this question requires marginalising the effect of weights, i.e., estimating $\int \tau(s \mid \boldsymbol{\lambda})\, d\boldsymbol{\lambda}$. In the absence of an analytical expression, this integral must be approximated. A standard MC method may not be ideal, as it requires thousands of samples, and for each sample, a simulation must be run across all budget configurations $N \times K$ – a process that could take months or even years to complete.

Instead, we estimate this using a Quasi-Monte Carlo (QMC) method with the Halton sequence [236]. QMC improves upon MC by replacing random sampling with low-discrepancy sequences, such as Halton, which are designed to cover the space more uniformly. These sequences minimise gaps and overlaps, leading to more accurate approximations with fewer samples, without sacrificing much in terms of accuracy.

To satisfy the constraint $\sum_d \lambda_d = 1$, we followed the method proposed by Smith *et al.* [237]. This technique generates sequences in $D-1$ dimensions, appends a value of 1, sorts the resulting list, and then computes the differences between adjacent values to derive the weights λ_d. This transformation effectively maps the original $D-1$ dimensional sequence onto a simplex — a geometric structure in $D-1$ dimensions where each point is D-dimensional and inherently satisfies the required weighted sum condition. See Figure 4.7 for an illustration of the weights.

In Figure 4.8, we present results from the DREsS dataset using 50 QMC-sampled weight vectors, independently generated for each N and K configuration, and strategy. Figure 4.9 shows the corresponding outcomes for the standard single-criterion BCJ approach. Across all configurations, BCJ was not outperformed by any other strategy. However, the combination of MCP with entropy-based pair selection consistently achieved the best performance among the multi-criteria variants.

The results for the BU dataset are presented in Figure 4.10 (multi-criteria variants) and Figure 4.11 (single-criterion BCJ), and they closely mirror the findings from the DREsS dataset. Notably, the combination of MCP and entropy pair selectors performed slightly better here than in the DREsS results. A closer analysis reveals that this MCP–entropy pairing consistently outperformed other multi-criteria variants. MCP with NRP also showed strong performance, though not to the same extent. As observed across all experiments, MCR combined with any pair selector consistently underperformed compared to the MCP ranking method.

In terms of robustness, BCJ emerges as the most effective method, with the MCP–entropy strategy following closely behind. This presents practitioners with a meaningful trade-off: BCJ offers a holistic and potentially faster decision-making process, requiring consideration of only a single comprehensive dimension. Conversely, multi-criteria approaches may take slightly longer, as each decision involves comparing all LOs within a pair. However, they offer richer insights into item discrimination, which can enhance feedback quality and improve transparency.

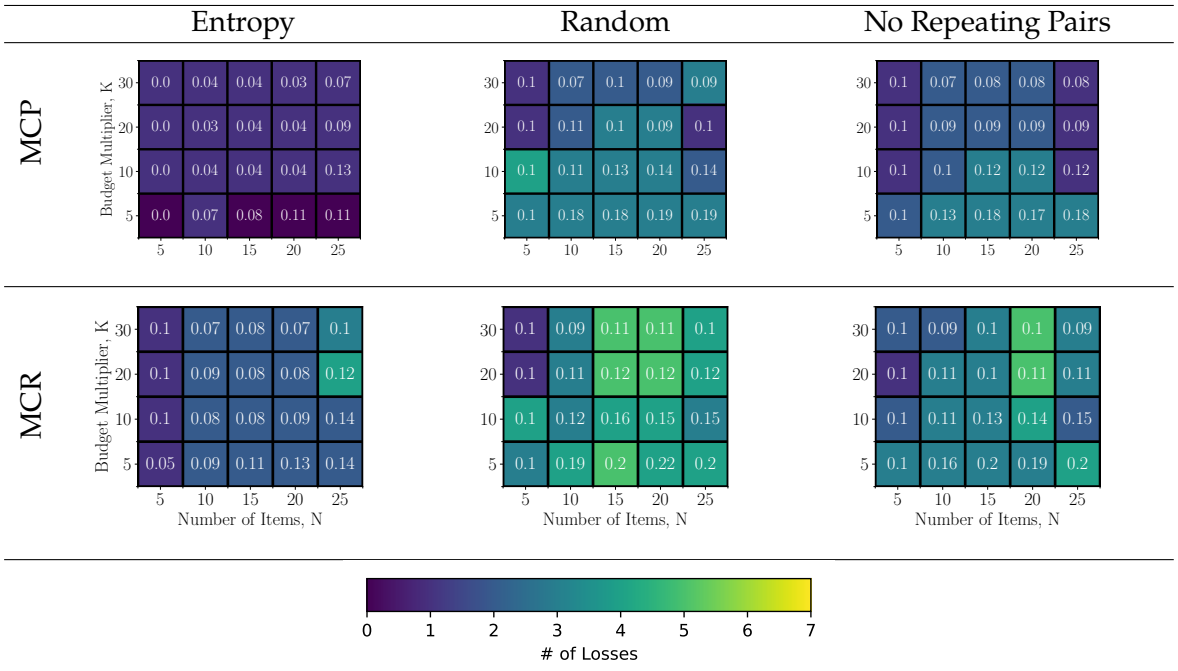


Figure 4.8: An illustration of the statistical comparison of multi-criteria strategies on the DREsS dataset using 50 QMC-sampled weight vectors. The plots show that, overall, the MCP ranking method significantly outperforms the MCR ranking method. Among all strategies, the combination of MCP with entropy-based pair selection achieves the best performance. This winning strategy is only outperformed by the standard BCJ approach in one configuration.

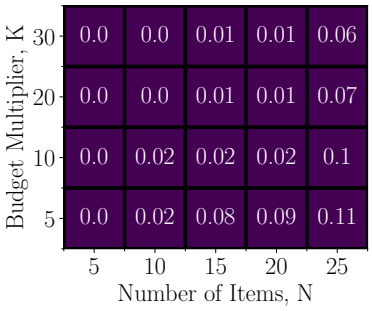


Figure 4.9: An illustration of the statistical comparison between the standard BCJ strategy and multi-criteria variants. Across all comparisons using 50 QMC-sampled weight vectors, the BCJ strategy with entropy-based pair selection was not dominated by any other method. The strategy combining MCP with entropy also performed strongly, but was consistently outperformed by BCJ in these configurations.

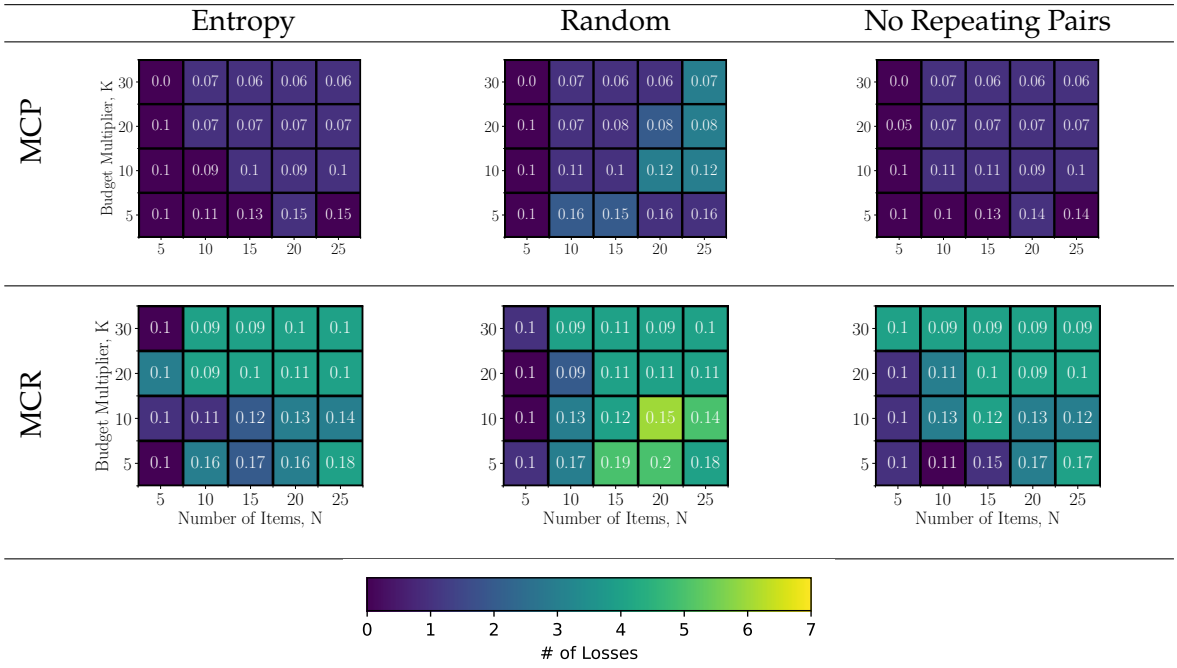


Figure 4.10: Statistical comparison of multi-criteria strategies for the BU dataset using QMC-sampled random weight vectors. The plots indicate that the MCP ranking method consistently outperformed the MCR approach. Among the MCP-based strategies, the pairing with the entropy selector showed a slight advantage over the combination with NRP for this dataset.

4.4.3 Reassessing Scale Separation Reliability as a Metric

Across all experiments, our findings challenge the reliability of SSR as a definitive metric for ranking accuracy in CJ. We observed several cases where the target ranking was achieved (i.e. a τ score of 0), even though the corresponding SSR score fell well below the commonly recommended threshold of 0.7. Conversely, higher SSR scores did not consistently lead to more accurate rankings. This suggests that SSR may be more reflective of the volume of comparisons — and the resulting confidence in those comparisons — rather than the actual quality of the final ranking.

One particularly revealing insight was the wide variation in SSR scores across experiments. For instance, the lowest SSR score observed was 0.27 for $N = 5$, $K = 5$, yet the τ score was 0.1. The highest SSR score, 0.92, occurred at $N = 25$, $K = 30$. Interestingly, in several cases where the τ score was 0 (indicating perfect ranking), the SSR score remained at 0.56 — below the recommended threshold — for both $N = 5$, $K = 5$ and $N = 5$, $K = 30$.

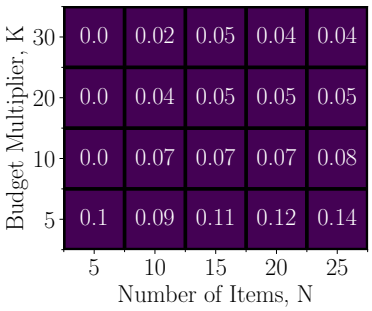


Figure 4.11: Statistical comparison of single-criterion BCJ combined with entropy-based pair selection versus multi-criteria variants, evaluated across 50 QMC-sampled weight vectors. The results show that BCJ with entropy selection was consistently competitive and not outperformed by any other method across the sampled weights.

In another instance, with $N = 5$, $K = 30$, the SSR score was again 0.56, but the τ score rose to 0.3. Across 38 experiments with an SSR score of 0.56, 25 achieved a τ score of 0.

For example, when $N = 5$ and $K = 10$, the average SSR score was 0.56, ranging from 0.47 to 0.56, yet 28 out of 50 runs resulted in a τ score of 0. Even at the lowest SSR score, the τ score was 0.1, and at the highest SSR score of 0.56, the τ score ranged from 0.0 to 0.1. These patterns suggest that SSR is more closely tied to the number of comparisons conducted, with higher SSR scores emerging from more extensive comparison sets. This raises important questions about the appropriateness of SSR as a measure of ranking accuracy in CJ, and highlights the need for further research into its role and limitations.

While an SSR score provides a general sense of overall agreement, it does not reveal where assessors specifically disagreed. In contrast, MAP and EAP offer visual insights into which item pairs are contributing to disagreement, enabling a more targeted understanding of reliability, as demonstrated in Section 3.5.1.

4.5 Conclusion

Bayesian active learning for CJ introduces a new paradigm for efficiently collecting data through pairwise comparisons to produce accurate rankings. However, key limitations remain — notably, the lack of mechanisms to quantify decision reliability and the reliance on holistic comparisons rather than multi-criteria evaluations.

This chapter addresses these gaps with two core methodological contributions:

- **MCP:** A robust method for aggregating criterion-level judgements that maintains overall ranking accuracy while preserving detailed performance data across individual criteria.
- **Active Learning Strategy:** An efficient approach that reduces uncertainty across all criteria simultaneously.

In educational assessment, our multi-criteria BCJ framework enables both detailed feedback on individual LOs and an overall student ranking. When MCP ranking is combined with entropy-based pair selection, performance is comparable to standard holistic BCJ, while offering richer diagnostic insights.

Experimental results show that MCP with entropy selection consistently performs well, often matching or surpassing the accuracy of holistic BCJ. Although holistic BCJ demonstrates greater robustness under marginalised criteria weights, our multi-criteria approach delivers fine-grained, actionable feedback with minimal impact on ranking accuracy. This makes it a valuable tool for formative assessment, allowing educators to understand student performance in depth without compromising reliability.

In summary, this chapter introduced and evaluated the MBCJ approach, extending the principles of BCJ to account for multiple assessment criteria simultaneously. The approach demonstrated how capturing richer dimensions of judgement can provide deeper insights into student performance and address some of the limitations of one-dimensional ranking. While the experiments highlighted both the potential and the challenges of MBCJ, they also raised important questions about how markers experience and interpret these different approaches in practice. Building on this, the next chapter shifts focus from technical development to empirical evaluation, comparing traditional marking, BCJ, and MBCJ in order to examine their relative transparency, usability, and trustworthiness from the perspective of those directly engaged in assessment.

Chapter 5

Rendering Transparency to Ranking in Educational Assessment

Having established the technical validity and educational potential of BCJ and MBCJ, this chapter focuses on their practical application in enhancing transparency in educational assessment. Transparency is crucial for building trust in assessment systems and ensuring that educators and students can understand and act upon the results. Through empirical evaluation involving professional markers, this chapter examines how BCJ and MBCJ impact perceptions of transparency, workload, and fairness compared to absolute marking methods.

5.1 Introduction

In UK HE, marking burdens mirror those in schools due to larger cohorts, diverse assessment formats, and intensified accountability, with consequences for staff wellbeing, consistency, and fairness [11, 13, 14, 80, 81, 82]. Post-COVID shifts and policy scrutiny have heightened demands for transparent, defensible assessment, yet human-judged tasks still suffer from variability despite automation and standardised rubrics [17, 125, 128, 129, 130, 85]. Traditional rubric grading and standard CJ either struggle with subjectivity or lack interpretability ("black-box") [238, 239, 147]. This chapter therefore examines BCJ and MBCJ, which incorporate priors, quantify uncertainty, and provide an auditable trail to target disagreements, support moderation, and promote equity [223, 240, 241, 242].

We apply BCJ/MBCJ to real HE assessments, demonstrating workload relief alongside improved reliability and transparency within sector governance frameworks [84].

The multi-criteria Bayesian approach (MBCJ) integrates multiple LOs into the CJ framework [240]. Traditional CJ operates on holistic judgement, collapsing distinct LOs into a holistic decision of which of a set of submissions being compared is best – and BCJ provides a Bayesian framework for this. By contrast, MBCJ allows each LO to be assessed independently, enabling the derivation of independent rankings aligned with each specific LO [240]. This nuanced approach provides an avenue for assessment of individual aspects, such as critical thinking, technical proficiency, and creativity, resulting in a more granular, detailed, assessment that mirrors rubric-based methods while maintaining the benefit of pairwise comparison structure of CJ. MBCJ also provides an overall holistic ranking that combines insight from each LO model, offering a balanced view across assessment criteria, like the rubric-based approach [240].

The MBCJ approach does not sacrifice depth for breadth; it retains the richness of qualitative insights inherent in CJ and expands them further. For example, if critical thinking and technical accuracy are weighted differently, the model can transparently reflect these nuances in the final ranking. This ability to disaggregate and re-aggregate rankings according to distinct LOs is particularly advantageous in high-stakes settings, where fair and transparent evaluation of multiple competencies is crucial, such as undergraduate final-year project dissertations. Through this process, MBCJ enhances transparency by clarifying the distinct and cumulative contributions of each LO to the final assessment, making the assessment process fairer and more comprehensible for all stakeholders [240].

For the first time, to illustrate the effectiveness of BCJ in real-world educational contexts, this paper employs a dataset from a UK HE setting to demonstrate the practical implementation of BCJ and MBCJ, and evaluate their impact on transparency, fairness, and accountability in assessment. By employing a combination of quantitative analyses and qualitative insights from professional markers, and views of experts in the field, we assess the efficacy of BCJ and MBCJ in providing a clear rationale for individual and aggregate judgements. Furthermore, we examine the usefulness of BCJ and MBCJ for identifying sources of ambiguity or conflict in assessments, particularly in contexts that demand high levels of transparency, such as national assessments [147]. 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.

Section 5.2 looks at the experiment's setup and the methodologies used; Section 5.3 looks at the results and discusses them; while in Section 5.4, we make our conclusions about the study.

5.2 Experimental Settings

In this chapter, 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. The official and ratified marks supplied by the module's lead lecturer serve as a ground-truth score for comparison. Three marking assistants who have experience of helping deliver the module perform traditional absolute marking, standard BCJ, and MBCJ. It should be noted that the submissions used were not originally marked by this cohort of assistants. Therefore, they encountered these submissions for the first time during the experiment conducted for this study. The process is discussed in more detail in Section 5.2.2.

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 Section 5.2.2. 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).

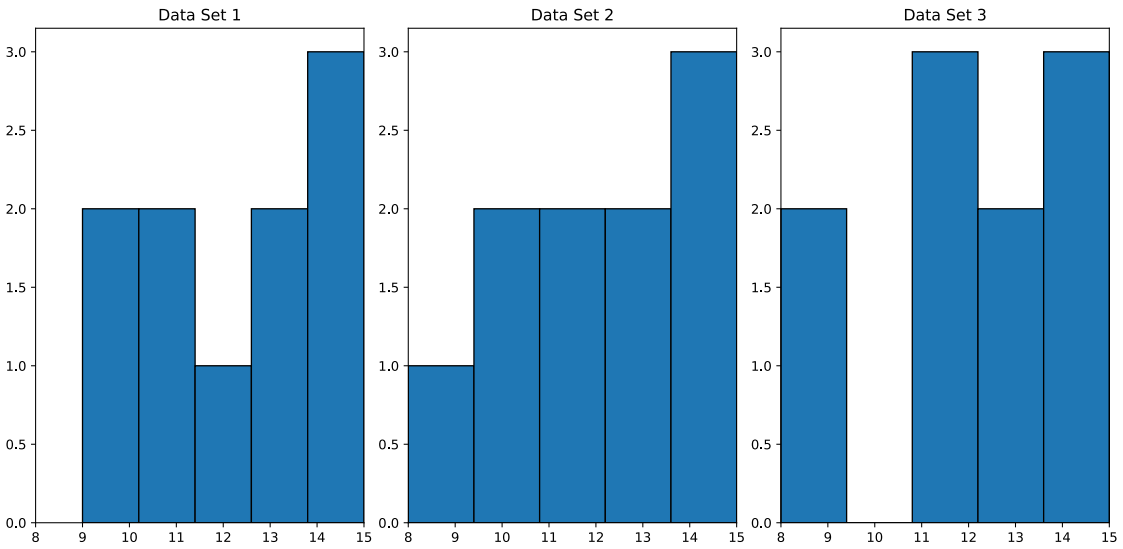


Figure 5.1: A histogram of marks for submissions in different groups: candidates for absolute marking, BCJ and MBCJ, entitled as dataset 1, 2 and 3 respectively. Clearly, the groups have similar distribution over the range between 8 and 15; this is important for a fair comparison between the groups.

5.2.1 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 5.2, 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 [239].

When assessors want to view the ranking of the items, they can view the results page, as demonstrated in Figure 5.3. 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

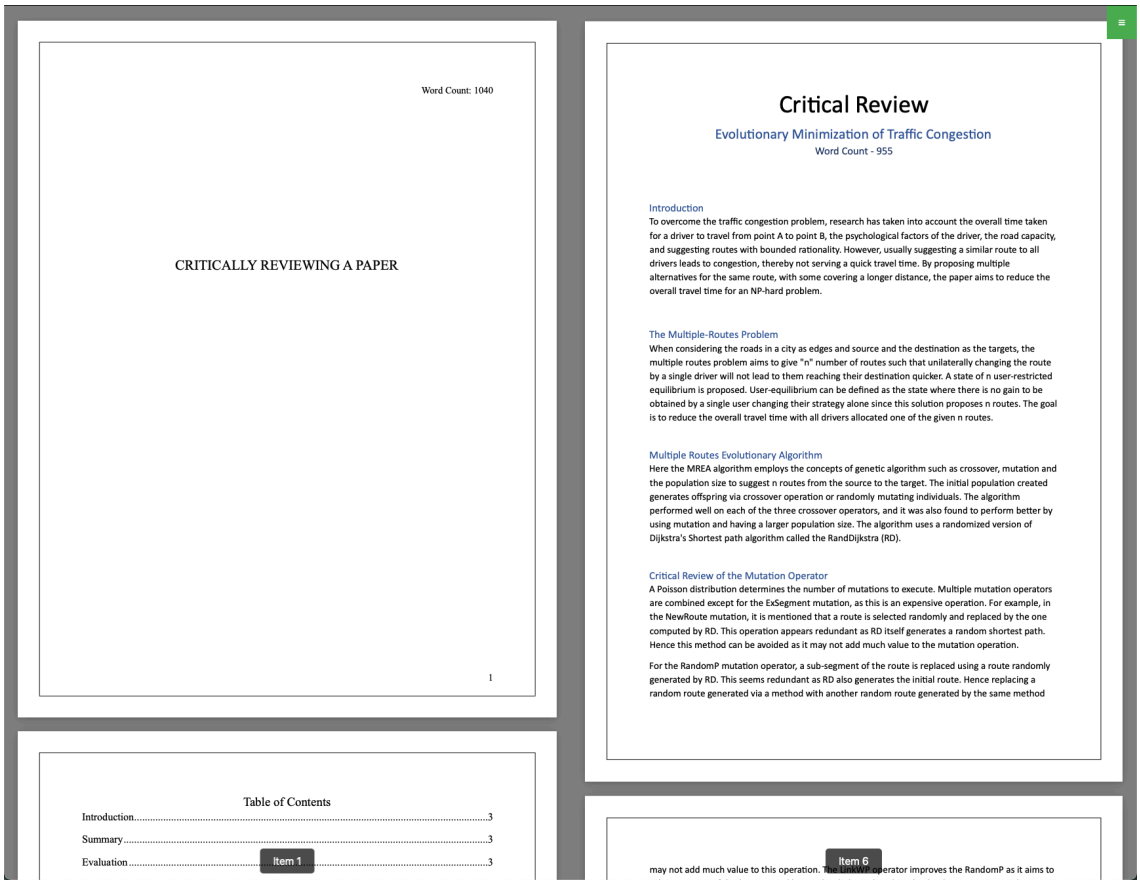


Figure 5.2: An example of the web app page for the standard BCJ's comparison. This page is what the assessor will see when they are making their judgements on the items being presented to them. Once they have pressed the corresponding button linked to the item they prefer, this will update the scores and then produce two new items for the assessor to compare.

have been compared. The ranks are calculated from the performance matrix using either BCJ or MBCJ while navigating to the results page.

Figure 5.4 shows the comparison screen that the assessors view when making their pair-wise comparisons. This screen is similar to the standard approach (see Figure 5.2), but has 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, 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

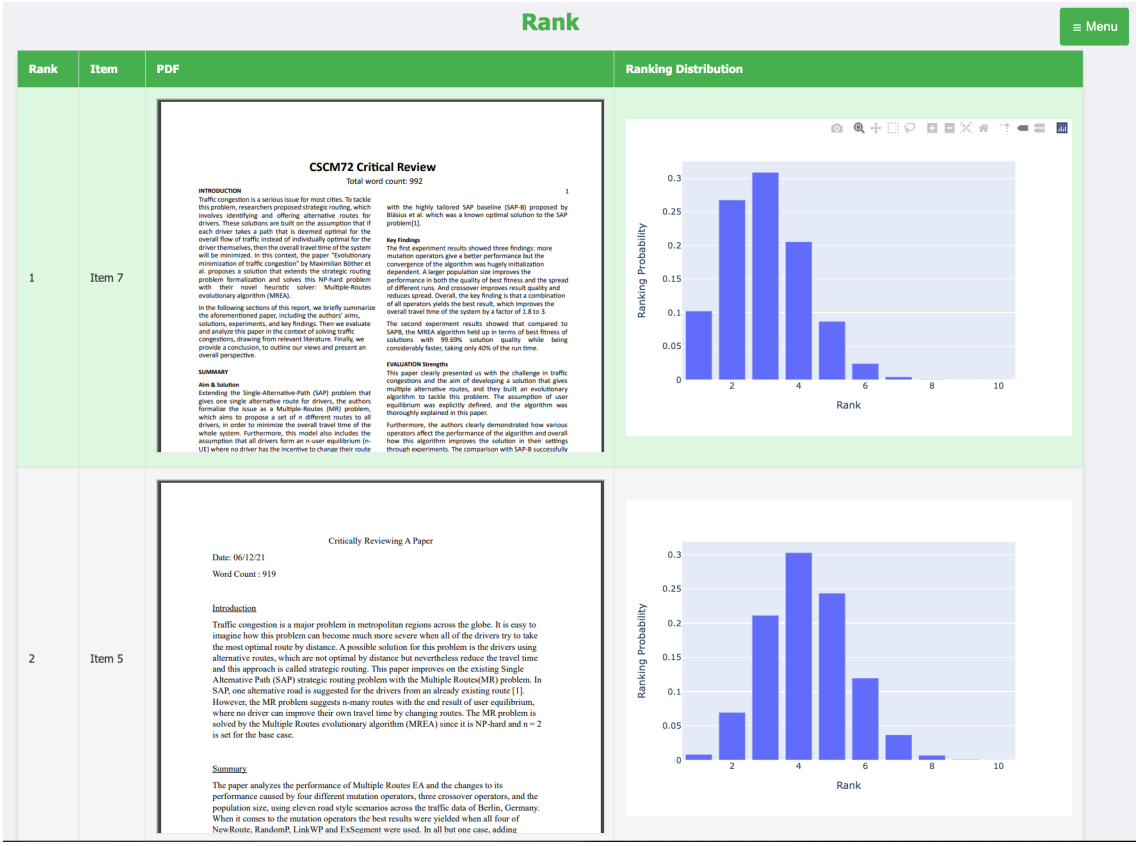


Figure 5.3: When the assessor wants to view the results, they can visit the results page. This web app page shows the items in order of their ranking, so the item ranked first will appear first on the page, and as the assessor scrolls down the page, they will then view the additional items until they 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.

preference matrices are updated, which enables the differential beta entropy to select the next pair of items to present to the assessor.

Figure 5.5 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 5.6). 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

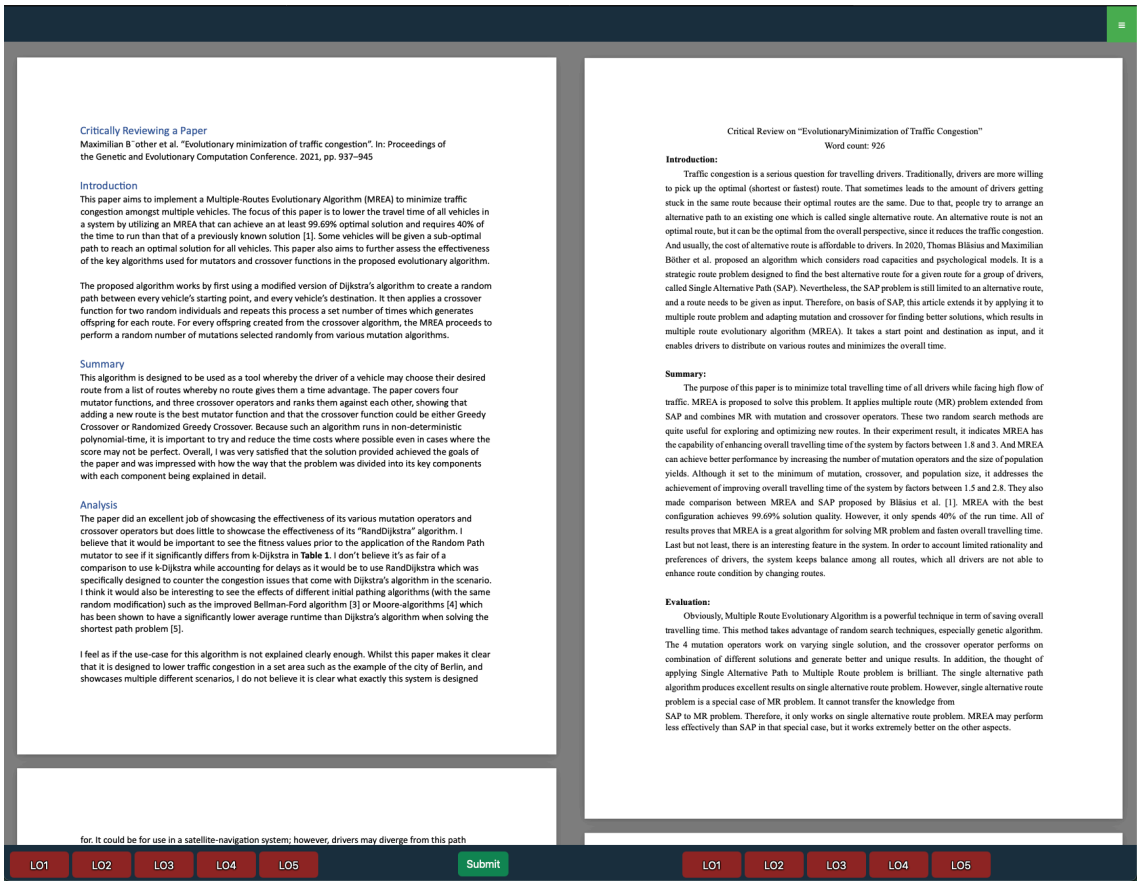


Figure 5.4: An example of the interface for the web app page for the multi-criteria comparison page. Like the standard BCJ page, this is where the assessor will make their decisions on the items displayed. However, they will need to make decisions based on individual LOs this time. They press the submit 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.

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.

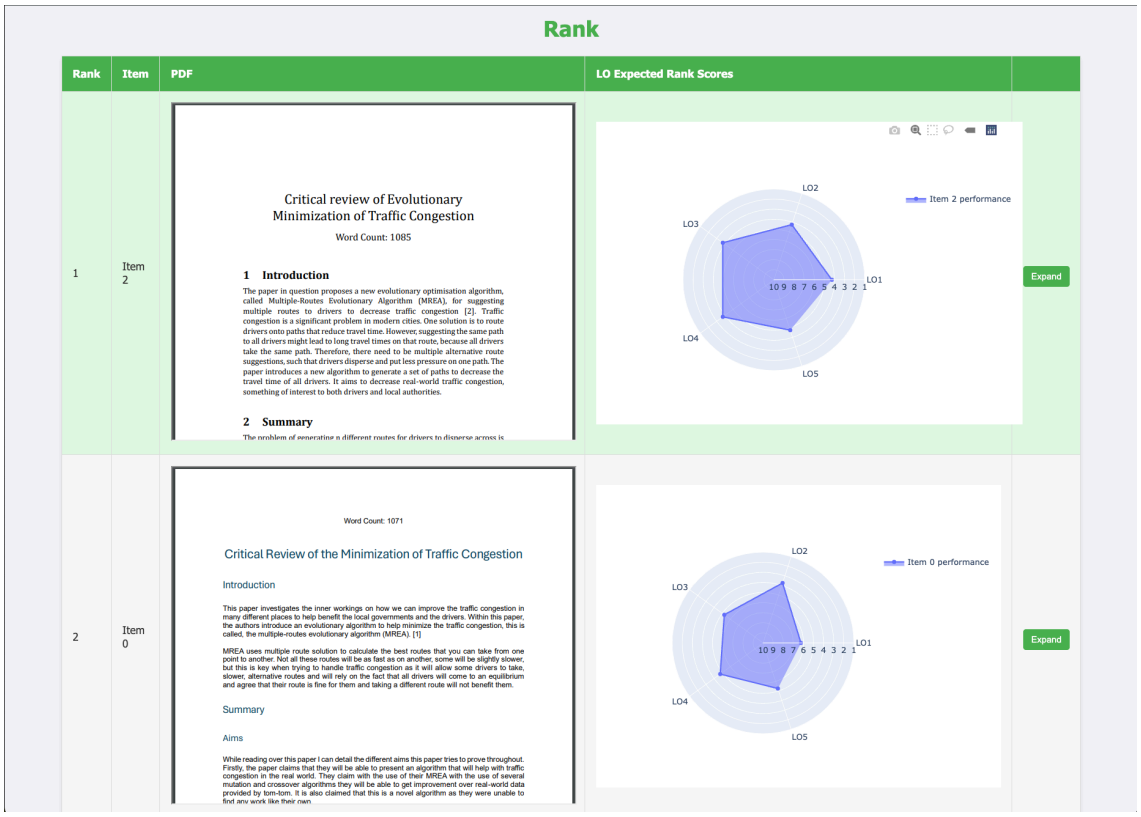


Figure 5.5: An example that presents the results of a multi-criteria BCJ web app showcasing transparency in expected rank E_r scores across different LOs using a radar plot for each item. An expand button is available for the assessor to be able to view the complete rank distributions for the individual LOs.

5.2.2 Research Approach

Three markers were recruited for this experiment. The markers were experienced educators 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 absolute marking, and a maximum of two hours 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).

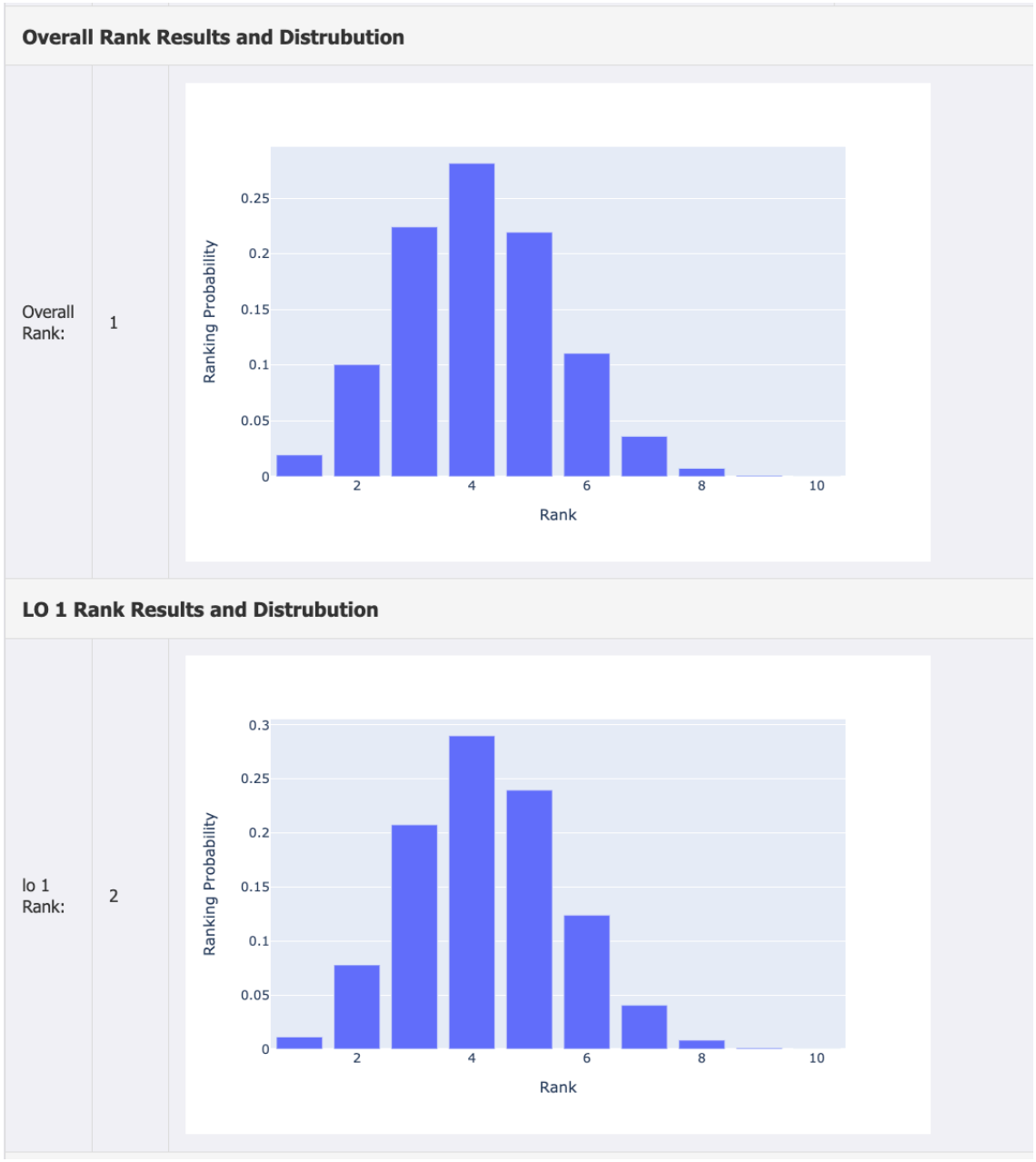


Figure 5.6: An example of the results of a multi-criteria BCJ web app showcasing transparency in ranking distributions across various LOs. The page provides an overview of the item's overall rank and distribution and its performance within each LO.

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).

5.3 Results and Discussion

In this section we look at the τ results against the target ranks, the markers' questionnaires, and the outcomes of the workshop session that took place with all the markers as well as expert interviews. We focus on if and how the BCJ and MBCJ are transparent and the issues raised to be mindful of when using BCJ and MBCJ.

While this study was conducted using HE students' work, existing literature demonstrates the effectiveness of CJ across educational levels, from primary to HE [243, 158, 162], showing its potential applicability beyond the HE context. These studies established CJ as a valid and efficient approach to assessment, particularly in subjects requiring holistic or qualitative evaluation.

Nonetheless, we acknowledge that the majority of the contextual framing draws on primary and secondary education. 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.

Thus, BCJ builds 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, 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.

5.3.1 τ Scores Against Target Rank

Marker one completed their absolute marking sample in two hours and 17 minutes (see Table 5.1). Compared to the Oracle's ranking, this produced a τ score of 0.3556 , marker two 0.4 and marker three 0.4.

Table 5.1: This summarises the performance of markers during the traditional absolute marking process. It also includes key metrics such as the total time spent by each marker, the number of pairwise comparisons conducted, and the corresponding τ scores.

Marker	Time	τ Score
1	2:17:03	0.3556
2	1:26:00	0.4
3	1:54:00	0.4

When we compare the markers with each other (see Table 5.2), marker one and two generated a τ score of 0.4, while marker one against marker three generated a τ score of 0.4, marker two and marker three generated a τ score of 0.4888. These scores show the markers were as far apart from each other in respect to their final ranks as they were from the target rank of the Oracle and, in the case of marker two and three, more so at 0.4888. This suggests that variation is a result of noise, as opposed to bias, in the process.

Table 5.2: The τ results of the final ranks created by the three markers when compared against each other for absolute marking. We can see that these compared to each other are not as close compared to the τ results from the Oracle's rank in Table 5.1, but we can see that marker one compared to marker two and marker one compared to marker three were the most similar with marker two and three being the furthest away.

Marker	Competitor	τ Score
1	2	0.4
1	3	0.4
2	3	0.4888

Performing BCJ, marker one completed 49 comparisons in one hour and seven minutes. Their τ score was 0.1556 (see Table 5.3), while marker two took two hours to complete their comparisons and completed a total of 46 with a τ score of 0.1778 and marker

three took one hour and twenty-two minutes and done a total of 50 comparisons and had a τ score of 0.1778.

All markers produced a better τ score with BCJ considering the Oracle's target rank. When comparing these scores to those from absolute marking methods, it becomes evident that the traditional approach was inconsistent. The marks awarded using absolute marking were highly subjective and varied significantly from one marker to another.

Table 5.3: This summarises the performance of markers during the BCJ process. It also includes key metrics such as the total time spent by each marker, the number of pairwise comparisons conducted, the resulting rank assigned based on their contributions, and the corresponding τ scores.

Marker	Time	No. of Comparisons	τ Score
1	1:07:00	49	0.1556
2	2:00:00	46	0.1778
3	1:22:00	50	0.1778
Combined			0.02

When we compare the markers against each other (see Table 5.4), marker one and two generated a τ score of 0.3333, in contrast, marker one against marker three generated a τ score of 0.2889, marker two and marker three generated a τ score of 0.2667, these results show that the markers become more aligned with each other compared to absolute marking.

While the markers are showing that they are more aligned with the Oracle's ranking when it comes to BCJ, interestingly, when comparing their generated final ranks, they are further away from each other regarding τ scores than they are compared to the Oracle's mark.

Table 5.4: The τ results of the final ranks created by the three markers when compared against each other for BCJ. We can see that these are not as close compared to each other as the τ results compared to the Oracle's rank in Table 5.3, but we can see that marker two and three were the most similar, with marker one and three the next closest.

Marker	Competitor	τ Score
1	2	0.3333
1	3	0.2889
2	3	0.2667

With MBCJ, marker one completed 37 comparisons in two hours. Their τ score was 0.1333 (see Table 5.5), while marker two took one hour and forty-nine minutes to complete their comparisons and did a total of 57 with a τ score of 0.1556 and marker three

took one hour and twenty-two minutes and done a total of 53 comparisons and had a τ score of 0.2886.

These results show that all markers were more closely aligned with the Oracle's target ranks when using MBCJ, compared to both traditional and BCJ marking methods—except for Marker Three. Although Marker Three exhibited a 62.3% increase in the τ score with MBCJ compared to BCJ, this still represented a 27.9% decrease relative to the traditional method. In contrast, Marker One demonstrated improvements of 14.3% and 62.5% over BCJ and traditional methods, respectively. Similarly, Marker Two showed gains of 12.5% and 56.2%. These are clearly consistent and substantial improvements.

Table 5.5: This summarises the performance of markers during the MBCJ process. It also includes key metrics such as the total time spent by each marker, the number of pairwise comparisons conducted, the resulting rank assigned based on their contributions, and the corresponding τ scores.

Marker	Time	No. of Comparisons	τ Score
1	2:00:00	37	0.1333
2	1:49:00	57	0.1556
3	1:22:00	53	0.2886
Combined			0.2

Overall, we can see that the results from the absolute marking approach reveal noticeable differences in both the time taken and the consistency of rank ordering among the markers. Marker one took the longest to complete the task, spending two hours and 17 minutes, and obtained a τ score of 0.3556, indicating a moderate level of alignment with the Oracle's ranking. In contrast, markers two and three completed their marking more quickly (one hour, twenty-six minutes, and one hour and fifty-four minutes, respectively) and achieved a slightly higher τ score of 0.4. This suggests a marginally better agreement with the target rank compared to marker one, despite the variation in time spent. The comparative analysis between markers shows that marker two and marker three were the

Table 5.6: The τ results of the final ranks created by the three markers when compared against each other for multi-criteria BCJ.

Marker	Competitor	τ Score
1	2	0.1111
1	3	0.2444
2	3	0.1778

least aligned with each other, yielding a τ score of 0.4888, highlighting greater inconsistency in rank ordering.

These findings indicate that absolute marking yielded the highest variability both in terms of time spent and rank consistency. In contrast, BCJ and MBCJ improved consistency across markers and alignment with the Oracle's ranking. MBCJ demonstrated the best overall performance, although individual differences among markers remained evident. This suggests that CJ methods, particularly MBCJ, could offer a more reliable alternative to traditional absolute marking by reducing subjective variability.

5.3.2 Questionnaire Results and Analysis

Marker one rated absolute 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 absolute 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 absolute 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 absolute 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 absolute marking still offered more direct feedback to students, which they felt could be integrated into MBCJ in the future.

Marker two found that absolute 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 absolute marking due to its transparency and ability to provide feedback. They believed BCJ and MBCJ were useful but would work best alongside absolute marking rather than replacing it. Ultimately, they had the most confidence in the

rankings generated by absolute marking, as it provided clear, section-by-section scores rather than relative comparisons between students.

Marker three found absolute 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 absolute 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 absolute 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 absolute marking, especially when multiple markers were involved, as it could help moderate inconsistencies. Ultimately, they had the most confidence in either absolute marking or MBCJ, with absolute 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 absolute 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 absolute 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, absolute marking remained the method in which markers placed the greatest trust, particularly regarding final marks and rankings.

Both absolute 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 absolute 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 absolute 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 absolute 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 absolute 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 absolute 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 absolute 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.

5.3.3 Workshop Results and Analysis

The three makers who took part in the experiment came together to discuss as a group about their experience 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. To which they all agreed they were.

The workshop began with a recap of the three marking methods: absolute marking, BCJ, and MBCJ. Participants were invited to reflect on their initial assumptions about these methods before reviewing their marking outcomes. Most expected absolute 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. absolute 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 absolute 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, absolute 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 absolute marking for feedback. This could reduce marking burden while maintaining transparency and student guidance.

Participants also reflected on how marking scales over larger cohorts. absolute 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 and inconsistencies would then be corrected by the BCJ system's ranking abilities. Which is an interesting point as in usual CJ implementations, this is how CJ is usually carried out, as it can be down with one or multiple markers contributing together [244]. 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 absolute 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 absolute 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 absolute 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 absolute marking in many contexts, provided feedback mechanisms were developed.

5.3.4 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 absolute 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 absolute 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.

5.3.4.1 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 absolute 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.

5.3.4.2 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 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 marking burden 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's 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.

5.3.4.3 Future Directions

A key theme was the shift in perception after reviewing BCJ and MBCJ results. Initially, absolute 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, absolute 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.

5.3.5 BCJ Transparency in the Assessment Procedure

Initially, absolute marking was perceived as the most transparent because it provided a structured, criteria-based approach with explicit marks and feedback by the markers. The markers felt that transparency came from clearly defined assessment rubrics, where each score was justified based on LOs. However, they also acknowledged that

absolute 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 absolute 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 5.3), 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 absolute marking.

MBCJ, however, was seen as a step towards greater transparency (see Figures 5.5 and 5.6). 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

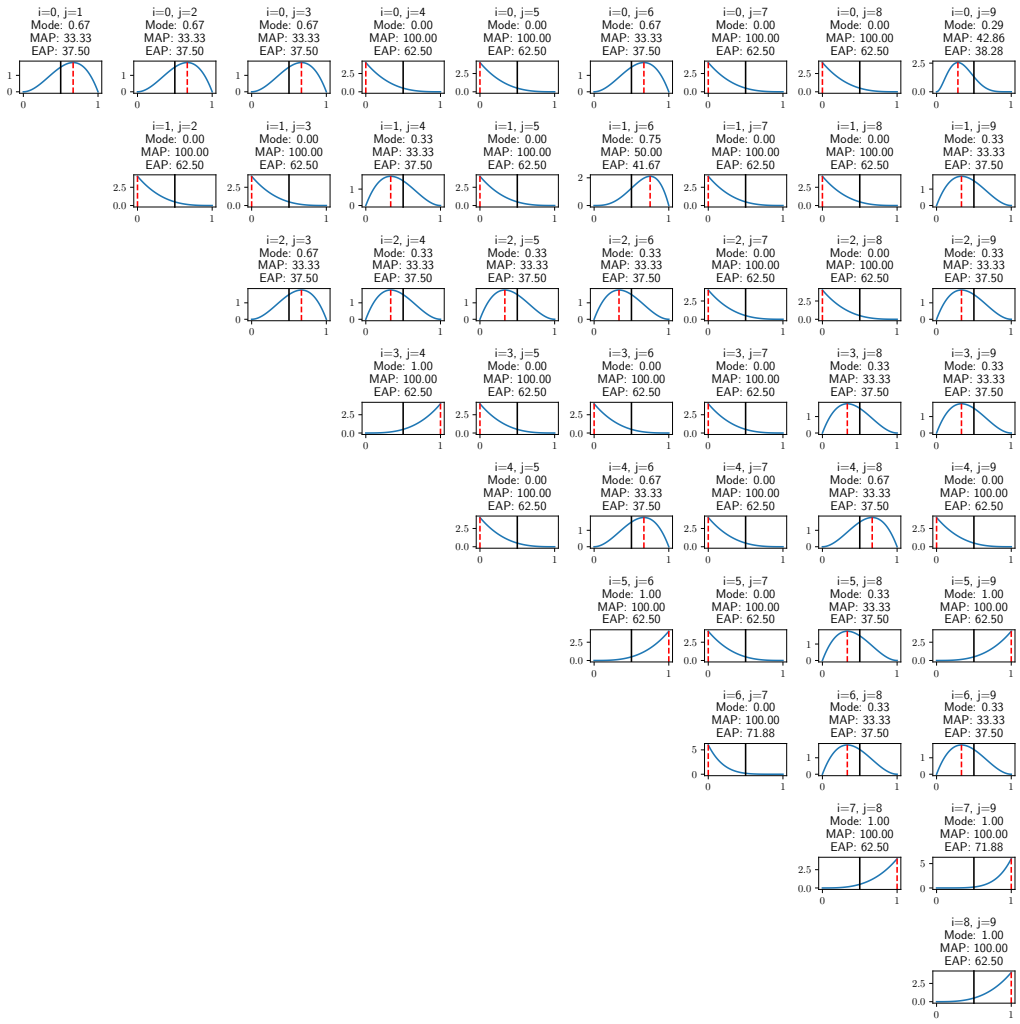


Figure 5.7: This shows the transparency in the decisions being made. The closer the distribution is to 0.5 (the black line), the more uncertain the markers are, meaning that half went one way and the other half went the other way. The red dotted line represents the mode β value from the decisions made, while the blue line shows the probability density function of the β distribution.

and fair alternative to traditional marking, but only if mechanisms for providing student feedback were integrated into the process.

Gray *et al.* [240] proposed new metrics to measure 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 $25^{\text{th}} - 75^{\text{th}}$ 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 5.7, 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 $25^{\text{th}} - 75^{\text{th}}$ quartile range, representing disagreements. For instance, in Figure 5.7, where $i = 0$ and $j = 6$, disagreements between the markers are evident, with a slight overall preference towards the item i with low agreement scores of an MAP of 33.3% and an EAP of 37.5%. We can additionally create a heatmap that produces these scores (see Figure 5.8) to visually identify the pairs that are dividing the crowd.

Expert two, when shown these outputs (Figures 5.7 and 5.8), 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 for 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 their 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 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

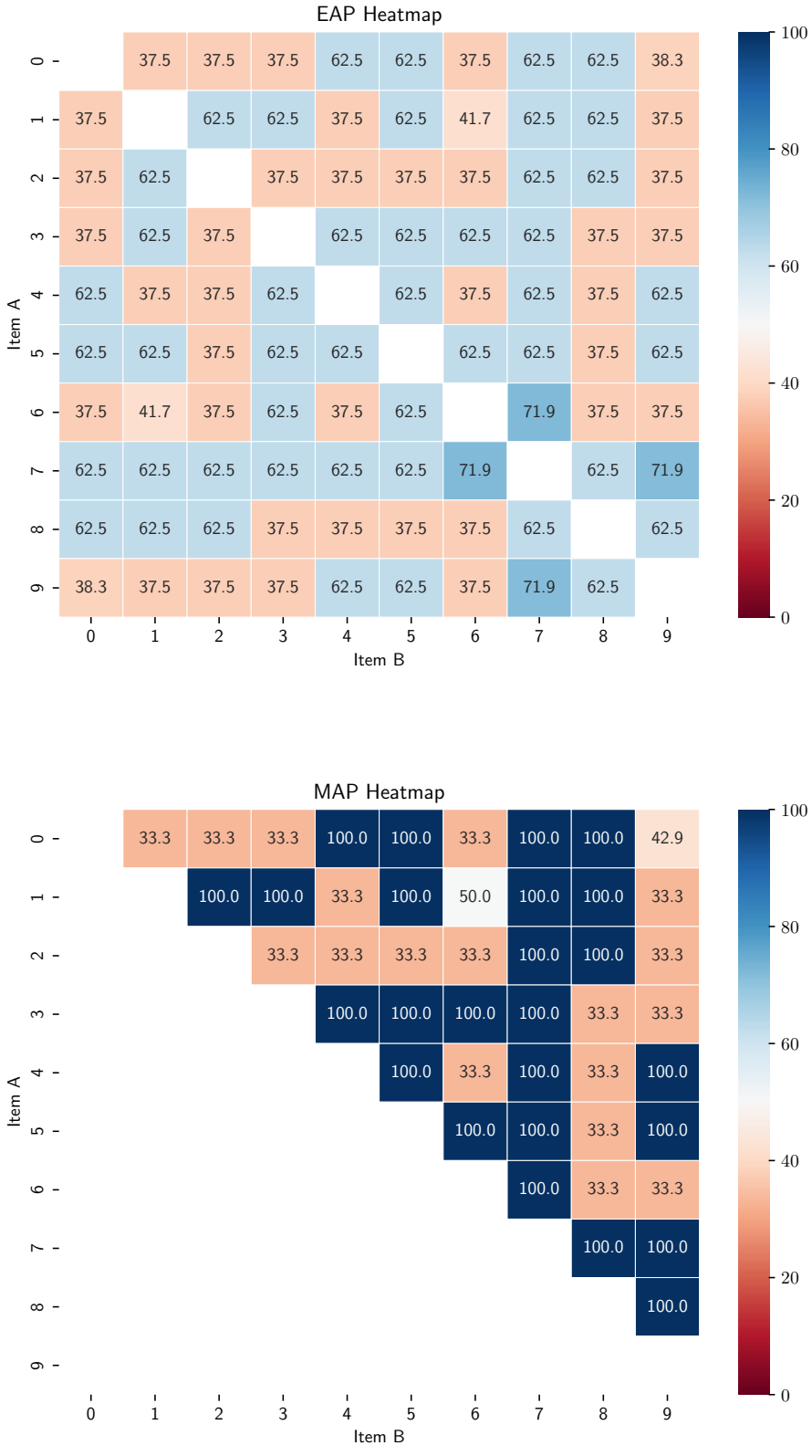


Figure 5.8: This shows an example of the EAP and MAP outputs. These heatmaps can be produced for all LOs for the MBCJ and holistically for the BCJ approach. Any value ≥ 50 indicates that the agreement is outside the 25^{th} and 75^{th} percentile ranges.

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 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 [242] 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.

5.3.6 Implementing BCJ

While the web apps used within this experiment are open-sourced and available on GitHub (see Section E and F 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.

In considering broader applicability, it is important to acknowledge several assumptions and contextual differences that may influence implementation across educational sectors. For example, assessment literacy among staff may vary significantly between primary, secondary, and HE settings, potentially affecting how CJ is understood and adopted. Institutional moderation practices also differ, with schools often operating under stricter external frameworks. Furthermore, student anonymity is a critical concern, particularly in school-based environments. While the BCJ/MBCJ system is a research

prototype, it is being actively developed with usability in mind. Although it has elements of 'pick-up-and-play' functionality, some familiarity with open-source software is beneficial. To support secure deployment, institutions would ideally host the web apps locally, enabling full control over data storage and ensuring student anonymity. Collaboration with IT departments to provide the necessary infrastructure will be key to successful implementation, especially in contexts where safeguarding and data governance are paramount.

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 2.3) – 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 [245].

If it is necessary to generate ranks after each comparison, a straightforward Monte Carlo (MC) version of the probabilistic computation for BCJ [223] can be used. In this case, with 300 items, the same machine can generate ranks within approximately 20 seconds. For MBCJ [240], 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 seconds [245] – 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 [246]. 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.

Additionally, while this study was conducted within a UK HE context, the foundational principles of CJ and BCJ, such as reducing subjectivity, improving transparency, and enhancing consistency, are relevant across all educational levels. It is important to acknowledge that the complexity of student work increases with educational level, meaning that CJ at primary level may be quicker to perform than at secondary, further education, or university level. Nonetheless, the core issues such as workload, assessment fairness, and moderation—remain consistent across sectors. The focus on secondary education in this chapter reflects the greater availability of statistical data in that context, but the implications of BCJ and MBCJ are potentially transferable to other levels. Future research should explore how these methods perform in school-based settings, particularly in relation to marking cultures, marking burden, and institutional moderation practices.

5.4 Conclusions

absolute 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.

For institutions or educators considering a pilot of MBCJ, several practical considerations should be addressed. These include establishing infrastructure for pairwise comparison (e.g. digital platforms), training staff in CJ principles, and integrating feedback mechanisms that align with existing assessment policies. Additionally, ensuring transparency in the selection of comparison pairs and providing students with interpretable outputs—such as radar plots or ranking distributions—can enhance trust and

engagement. These steps are essential to support successful implementation and scalability across diverse educational settings.

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.

The findings from these transparency-focused studies provide strong evidence of the practical value of BCJ and MBCJ in enhancing fairness, usability, and trust in educational assessment. Having demonstrated the technical, practical, and perceptual benefits of these frameworks across multiple studies, the next chapter concludes the thesis by synthesising these insights and reflecting on their broader implications for transforming assessment practices.

Chapter 6 also outlines clear directions for future research, bringing together the theoretical foundations, technical innovations, and empirical validations explored throughout this thesis to highlight how BCJ and MBCJ can support a fairer, more transparent, and efficient assessment landscape in education.

Chapter 6

Conclusions and Future Work

Marking and assessment play a critical role in education, yet traditional methods are often time-consuming and susceptible to inconsistencies. A significant challenge in assessment is the difficulty of evaluating absolute quality, which can lead to variability in grading outcomes. With the emergence of generative AI tools in education, new challenges and opportunities are arising for assessment practices. CJ has emerged as a promising alternative, addressing many issues associated with traditional marking. Still, it also introduces its own challenges, particularly in determining optimal comparison counts and maintaining ranking stability.

This research has demonstrated that BCJ offers substantial improvements over standard CJ methods. The introduction of BCJ mitigates the rank deterioration issue observed in traditional CJ, as accuracy improves with additional comparisons. Furthermore, the transparency of assessment processes remains a critical concern. While CJ enhances reliability in some respects, its opacity regarding rank generation can limit its acceptance as a direct replacement for marking. In response, this study has developed BCJ with transparency at its core, enabling users to understand how rankings are derived, allowing for input into the grade determination process, and providing insight into model predictions. This increased transparency makes BCJ a more viable alternative for educational assessment.

Expanding upon BCJ, this study has also introduced an MBCJ approach, which further refines assessment by evaluating student performance across individual learning objectives (LOs) while maintaining an overall ranking. This multi-criterion approach retains the benefits of standard BCJ while providing a more granular and informative assessment of student work. By integrating a weighted ensemble machine learning strategy

with differential entropy pair-picking, MBCJ achieves comparable ranking performance to single-criterion BCJ while offering greater transparency and deeper insights into student strengths and weaknesses.

Despite these advancements, challenges remain in scaling BCJ and MBCJ for broader educational use. The study has highlighted key considerations, such as the optimal number of comparisons required for stable rankings and the potential limitations of traditional accuracy metrics like SSR in evaluating CJ-based models. Additionally, while BCJ and MBCJ have demonstrated enhanced transparency and fairness, further research is needed to refine feedback mechanisms, ensuring that students receive not only their rankings but also meaningful insights into their performance and areas for improvement.

The findings of this research suggest that while traditional marking remains the most widely trusted method, it is also highly demanding and prone to inconsistencies. In contrast, BCJ and MBCJ provide structured assessment methods that reduce subjectivity, improve reliability, and align more closely with rubric-based marking practices. Educators found MBCJ particularly valuable for preserving the depth of feedback typically associated with rubrics while benefiting from the efficiency of CJ.

Future research will focus on integrating BCJ and MBCJ into real-world educational settings, exploring their scalability for large-scale assessments, and further developing feedback mechanisms to enhance student learning outcomes. Additionally, the potential applications of BCJ in high-stakes assessments and its adaptability to AI-driven educational tools warrant further investigation. By addressing these considerations, BCJ and MBCJ can be refined into robust, scalable, and transparent assessment methodologies that improve fairness, reliability, and efficiency in educational evaluation.

Ultimately, this study supports the notion that structured CJ methods, particularly BCJ and MBCJ, have the potential to transform assessment practices by making them more transparent, fair, and informative while reducing the marking burden on assessors. However, further refinements and adaptations will be necessary to facilitate widespread adoption within classroom settings and beyond.

This thesis has explored the challenges and limitations of traditional assessment methods in education, particularly their susceptibility to inconsistency and subjectivity. Marking and assessment play a critical role in education, yet traditional methods are often time-consuming and susceptible to inconsistencies. A significant challenge in assessment is the difficulty of evaluating absolute quality, which can lead to variability in grading

outcomes. With the emergence of generative AI tools in education, new challenges and opportunities are arising for assessment practices. CJ has emerged as a promising alternative, addressing many issues associated with traditional marking. Still, it also introduces its own challenges, particularly in determining optimal comparison counts and maintaining ranking stability.

CJ has emerged as a promising alternative to conventional marking, addressing some of these challenges by leveraging paired comparisons to establish rank order. However, traditional CJ methods encounter issues related to model stability and transparency, particularly when the number of comparisons grows. This study has demonstrated that standard CJ methods, particularly those using the BTM, struggle when scaling beyond a certain threshold, leading to deteriorating rank accuracy. Our findings indicate that while a recommended minimum of comparisons is suggested, higher values are required for improved reliability as the number of comparisons increases.

This research has demonstrated that BCJ offers substantial improvements over standard CJ methods. The introduction of BCJ mitigates the rank deterioration issue observed in traditional CJ, as accuracy improves with additional comparisons. Furthermore, the transparency of assessment processes remains a critical concern. While CJ enhances reliability in some respects, its opacity regarding rank generation can limit its acceptance as a direct replacement for marking. In response, this study has developed BCJ with transparency at its core, enabling users to understand how rankings are derived, allowing for input into the grade determination process, and providing insight into model predictions. This increased transparency makes BCJ a more viable alternative for educational assessment.

Expanding upon BCJ, this study has also introduced an MBCJ approach, which further refines assessment by evaluating student performance across individual LOs while maintaining an overall ranking. This multi-criterion approach retains the benefits of standard BCJ while providing a more granular and informative assessment of student work. By integrating a weighted ensemble machine learning strategy with differential entropy pair-picking, MBCJ achieves comparable ranking performance to single-criterion BCJ while offering greater transparency and deeper insights into student strengths and weaknesses.

Despite these advancements, challenges remain in scaling BCJ and MBCJ for broader educational use. The study has highlighted key considerations, such as the optimal number of comparisons required for stable rankings and the potential limitations of traditional accuracy metrics like SSR in evaluating CJ-based models. Additionally, while BCJ and

MBCJ have demonstrated enhanced transparency and fairness, further research is needed to refine feedback mechanisms, ensuring that students receive not only their rankings but also meaningful insights into their performance and areas for improvement.

Through real-world application and empirical evaluation, the research demonstrated that BCJ and MBCJ improve the reliability, consistency, and transparency of student ranking compared to traditional marking and standard CJ. The findings of this research suggest that while traditional marking remains the most widely trusted method, it is also highly demanding and prone to inconsistencies. In contrast, BCJ and MBCJ provide structured assessment methods that reduce subjectivity, improve reliability, and align more closely with rubric-based marking practices. Educators found MBCJ particularly valuable for preserving the depth of feedback typically associated with rubrics while benefiting from the efficiency of CJ. However, the study also identified the need for further refinement, particularly in integrating detailed feedback mechanisms to support student learning and making these approaches more accessible for large-scale adoption.

Future research will focus on integrating BCJ and MBCJ into real-world educational settings, exploring their scalability for large-scale assessments, and further developing feedback mechanisms to enhance student learning outcomes. Additionally, the potential applications of BCJ in high-stakes assessments and its adaptability to AI-driven educational tools warrant further investigation. By addressing these considerations, BCJ and MBCJ can be refined into robust, scalable, and transparent assessment methodologies that improve fairness, reliability, and efficiency in educational evaluation.

It should be noted that both BCJ and multi-criteria BCJ are generalisable to a range of applications. For instance, the work of [170] applied Comparative Judgement using a Bayesian Bradley–Terry model to evaluate classifiers based on a single accuracy criterion. A direct comparison with BCJ could be performed in this context.

This study supports the notion that structured CJ methods, particularly BCJ and MBCJ, have the potential to transform assessment practices by making them more transparent, fair, and informative while reducing the marking burden on assessors. However, further refinements and adaptations will be necessary to facilitate widespread adoption within classroom settings and beyond.

6.1 Future Work

A common theme throughout this research is that assessment is an important element in education. However, feedback is essential to enable students to grow and develop. Therefore, future work on this research would be to expand the proposed BCJ and MBCJ approaches to allow feedback to be presented, while ensuring that the core fundamentals of CJ, which reduce marking burden, are not compromised or lost. Additionally, future work would be to build on the open-sourced software packages we have made available to make this accessible to even more people. To truly help disrupt and revolutionise the marking experience for all assessors.

6.2 Final Reflections

This research began with a simple but persistent question: how can we assess student work in a way that is both rigorous and human-centred? In answering this, the thesis has shown that it is possible to combine probabilistic modelling, machine learning, and educational theory to build assessment systems that are not only statistically sound but also pedagogically meaningful. Bayesian Comparative Judgement and its multi-criteria extension offer educators new tools—ones that reduce marking load, increase fairness, and make the ranking process more transparent. Yet these tools are not just algorithms; they are interventions in a long-standing challenge of education: how to balance efficiency, trust, and feedback. This thesis has taken steps toward resolving that tension, and it is my hope that these contributions will serve not only as a technical foundation but as a conceptual shift in how we approach the design and delivery of assessment in the years ahead.

Bibliography

[1] P. Finn and R. Cinpoes, "The impact of COVID-19 on A-Levels since 2020, and what it means for higher education in 2022/23," https://blogs.lse.ac.uk/politicsandpolicy/impact-of-covid19-on-a-levels, 2022.

[2] I. Nisbet and S. Shaw, *Is Assessment Fair?* Sage, 2020.

[3] B. Jeffreys, "A-levels: Students told most will get first-choice university place," https://www.bbc.co.uk/news/education-62518040, 2022.

[4] N. Everett and J. Papageorgiou, *Investigating the Accuracy of Predicted: A Level Grades as Part of 2009 UCAS Admission Process*. Department for Business Innovation & Skills, 2011.

[5] R. Watermeyer, T. Crick, C. Knight, and J. Goodall, "COVID-19 and digital disruption in UK universities: afflictions and affordances of emergency online migration," *Higher Education*, vol. 81, pp. 623–641, 2021.

[6] T. Crick, C. Knight, R. Watermeyer, and J. Goodall, "The Impact of COVID-19 and "Emergency Remote Teaching" on the UK Computer Science Education Community," in *Proceedings of UK and Ireland Computing Education Research Conference (UKICER'20)*, 2020.

[7] E. Marchant, C. Todd, M. James, T. Crick, R. Dwyer, and S. Brophy, "Primary school staff perspectives of school closures due to COVID-19, experiences of schools reopening and recommendations for the future: a qualitative survey in Wales," *PLOS ONE*, vol. 16, no. 12, p. e0260396, 2021.

[8] T. Crick, C. Knight, R. Watermeyer, and J. Goodall, "The International Impact of COVID-19 and "Emergency Remote Teaching" on Computer Science Education Practitioners," in *Proceedings of IEEE Global Engineering Education Conference (EDUCON'21)*, 2021, pp. 1048–1055.

[9] A. Siegel, M. Zarb, B. Alshaigy, J. Blanchard, T. Crick, R. Glassey, J. R. Holt, C. Latulipe, C. Riedesel, M. Senapathi, Simon, and D. Williams, "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)*, 2021.

[10] E. Lowthian, H. Abbasizanjani, S. Bedston, A. Akbari, L. Cowley, R. Fry, R. K. Owen, J. Hollinghurst, I. Rudan, J. Beggs, E. Marchant, F. Torabi, S. de Lusignan, T. Crick, G. Moore, A. Sheikh, and R. A. Lyons, "Trends in SARS-CoV-2 infection and vaccination in school staff, students, and their household members from 2020-2022 in Wales, UK: an electronic cohort study," *Journal of the Royal Society of Medicine*, 2023.

[11] R. Watermeyer, K. Shankar, T. Crick, C. Knight, F. McGaughey, J. Hardman, V. Suri, R. Chung, and D. Phelan, "'Pandemia': A reckoning of UK universities' corporate response to COVID-19 and its academic fallout," *British Journal of Sociology of Education*, vol. 42, no. 5-6, pp. 651–666, 2021.

[12] K. Shankar, D. Phelan, V. Suri, R. Watermeyer, C. Knight, and T. Crick, ""The COVID-19 Crisis is Not the Core Problem": Experiences, Challenges, and Concerns of Irish Academia in the Pandemic," *Irish Educational Studies*, vol. 40, no. 2, pp. 169–175, 2021.

[13] F. McGaughey, R. Watermeyer, K. Shankar, V. Suri, C. Knight, T. Crick, J. Hardman, D. Phelan, and R. Chung, "'This can't be the new norm': academics' perspectives on the COVID-19 crisis for the Australian University Sector," *Higher Education Research & Development*, vol. 41, no. 7, 2022.

[14] J. Hardman, R. Watermeyer, K. Shankar, V. Suri, T. Crick, C. Knight, F. McGaughey, and R. Chung, ""Does anyone even notice us?" COVID-19's impact on academics' well-being in a developing country," *South African Journal of Higher Education*, vol. 36, no. 1, pp. 1–19, 2022.

[15] T. Crick, "COVID-19 and Digital Education: A Catalyst for Change?" *ITNOW*, vol. 63, no. 1, 2021.

[16] R. Ward, O. Phillips, D. Bowers, T. Crick, J. H. Davenport, P. Hanna, A. Hayes, A. Irons, and T. Prickett, "Towards a 21st Century Personalised Learning Skills Taxonomy," in *Proceedings of IEEE Global Engineering Education Conference (EDUCON'21)*, 2021, pp. 344–354.

[17] R. Watermeyer, T. Crick, and C. Knight, "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*, vol. 27, no. 2, pp. 148–162, 2022.

[18] A. Irons and T. Crick, "Cybersecurity in the Digital Classroom: Implications for Emerging Policy, Pedagogy and Practice," in *Higher Education in a Post-COVID World: New Approaches and Technologies for Teaching and Learning*. Emerald Publishing, 2022, pp. 231–244.

[19] T. Crick, C. Knight, and R. Watermeyer, "Reflections on a Global Pandemic: Capturing the Impact of COVID-19 on the UK Computer Science Education Community," in *Proceedings of UK and Ireland Computing Education Research Conference (UKICER'22)*, 2022.

[20] E. Thomas, T. Crick, and G. Beauchamp, "Envisioning the Post-COVID "New Normal" for Education in Wales," *Wales Journal of Education*, vol. 25, no. 2, 2023.

[21] R. Watermeyer, T. Crick, and C. Knight, "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*, vol. 27, no. 2, pp. 148–162, 2022.

[22] T. Crick, T. Prickett, and J. Bradnum, "Exploring Learner Resilience and Performance of First-Year Computer Science Undergraduate Students during the COVID-19 Pandemic," in *Proceedings of 27th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE'22)*, 2022, pp. 519–525.

[23] R. Ward, T. Crick, J. H. Davenport, P. Hanna, A. Hayes, A. Irons, K. Miller, F. Moller, T. Prickett, and J. Walters, "Using skills profiling to enable badges and micro-credentials to be incorporated into higher education courses," *Journal of Interactive Media in Education*, vol. 2023(1), no. 10, pp. 1317–1336, 2023.

[24] C. Knight, C. Conn, T. Crick, and S. Brooks, "Divergences in the Framing of Inclusive Education across the UK: A Four Nations Critical Policy Analysis," *Educational Review*, 2023.

[25] S. Weale, "A-level results day will not be 'pain-free', head of UCAS says," https://www.theguardian.com/education/2022/aug/15/a-level-results-day-not-pain-free-head-of-ucas-says, 2022.

[26] R. Luckin, W. Holmes, M. Griffiths, and L. Forcier, "Intelligence Unleashed: An argument for AI in Education," Pearson Education, Tech. Rep., 2016.

[27] A. Namoun and A. Alshanqiti, "Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review," *Applied Sciences*, vol. 11, no. 1, p. 237, 2020.

[28] J. L. Rastrollo-Guerrero, J. A. Gómez-Pulido, and A. Durán-Domínguez, "Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review," *Applied Sciences*, vol. 10, no. 3, p. 1042, 2020.

[29] Y. K. Dwivedi, L. Hughes, E. Ismagilova, G. Aarts, C. Coombs, T. Crick, Y. Duan, R. Dwivedi, J. Edwards, A. Eirug, V. Galanos, P. V. Ilavarasan, M. Janssen, P. Jones, A. Kumar Kar, H. Kizgin, B. Kronemann, B. Lal, B. Lucini, R. Medaglia, K. Le Meunier-FitzHugh, L. Le Meunier-FitzHugh, S. Misra, E. Mogaji, S. Sharma, J. Bahadur Singh, V. Raghavan, R. Raman, N. Rana, S. Samothrakis, J. Spencer, K. Tamilmani, A. Tubadji, P. Walton, and M. Williams, "Artificial Intelligence (AI): Multidisciplinary Perspectives on Emerging Challenges, Opportunities, and Agenda for Research, Practice and Policy," *International Journal of Information Management*, vol. 53, no. 101994, 2021.

[30] D. A. Shafiq, M. Marjani, R. A. A. Habeeb, and D. Asirvatham, "Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review," *IEEE Access*, 2022.

[31] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala, "Predicting Student Performance Using Personalized Analytics," *Computer*, vol. 49, no. 4, pp. 61–69, 2016.

[32] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 11, 2022.

[33] Z. Iqbal, J. Qadir, A. Noor Mian, and F. Kamiran, "Machine Learning Based Student Grade Prediction: A Case Study," *arXiv*, 2017.

[34] V. Vijayalakshmi and K. Venkatachalapathy, "Comparison of Predicting Student's Performance using Machine Learning Algorithms," *International Journal of Intelligent Systems and Applications*, vol. 12, pp. 34–45, 2019.

[35] B. Yousafzai, M. Hayat, and S. Afzal, "Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student," *Education and Information Technologies*, vol. 25, pp. 4677–4697, 2020.

[36] S. Slade and P. Prinsloo, "Learning Analytics: Ethical Issues and Dilemmas," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1510–1529, 2013.

[37] B. Williamson, S. Bayne, and S. Shay, "The datafication of teaching in Higher Education: critical issues and perspectives," *Teaching in Higher Education*, vol. 25, no. 4, pp. 351–365, 2020.

[38] S. Akgun and C. Greenhow, "Artificial intelligence in education: Addressing ethical challenges in K-12 settings," *AI and Ethics*, vol. 2, pp. 431–440, 2019.

[39] B. Williamson, R. Eynon, and J. Potter, "Pandemic politics, pedagogies and practices: digital technologies and distance education during the coronavirus emergency," *Learning, Media and Technology*, vol. 45, no. 2, pp. 107–114, 2020.

[40] UK Public General Acts, "Education act 1988," 1988.

[41] D. Hutchison and I. Schagen, *How reliable is National Curriculum assessment?* NFER, 1994.

[42] J. Dillon and M. Maguire, *Becoming a teacher: Issues in secondary education*. McGraw-Hill Education (UK), 2011.

[43] J. Wellington, *Secondary education: The key concepts*. Routledge, 2007.

[44] P. Black and D. Wiliam, *Inside the black box: Raising standards through classroom assessment*. Granada Learning, 1998.

[45] S. Courtney, "Head teachers' experiences of school inspection under Ofsted's january 2012 framework," *Management in Education*, vol. 27, pp. 164 – 169, 2013.

[46] R. Woore, L. Molway, and E. Macaro, "Keeping sight of the big picture: a critical response to Ofsted's 2021 curriculum research review for languages," *The Language Learning Journal*, vol. 50, pp. 146 – 155, 2022.

[47] S. Earle and J. Turner, "What has happened to teacher assessment of science in english primary schools? revisiting evidence from the primary science quality mark," *Research in Science & Technological Education*, vol. 41, pp. 22 – 38, 2020.

[48] C. Beaumont, M. O'Doherty, and L. Shannon, "Reconceptualising assessment feedback: a key to improving student learning?" *Studies in Higher Education*, vol. 36, pp. 671 – 687, 2011.

[49] B. Apter, F. Sulla, and J. Swinson, "A review of recent large-scale systematic uk classroom observations, method and findings, utility and impact," *Educational Psychology in Practice*, vol. 36, pp. 367 – 385, 2020.

[50] K. Kerr, "Exploring student perceptions of verbal feedback," *Research Papers in Education*, vol. 32, pp. 444 – 462, 2017.

[51] J. Kulik and C.-L. C. Kulik, "Timing of feedback and verbal learning," *Review of Educational Research*, vol. 58, pp. 79 – 97, 1988.

[52] Y. Zhang, B.-L. Chen, J. Ge, C.-Y. Hung, and L. Mei, "When is the best time to use rubrics in flipped learning? a study on students' learning achievement, metacognitive awareness, and cognitive load," *Interactive Learning Environments*, vol. 27, pp. 1207 – 1221, 2018.

[53] R. Duran, A. Zavgorodniaia, and J. Sorva, "Cognitive load theory in computing education research: A review," *ACM Transactions on Computing Education (TOCE)*, vol. 22, pp. 1 – 27, 2022.

[54] A. P. de Moira, C. Massey, J. Baird, and M. Morrissy, "Marking consistency over time," *Research in Education*, vol. 67, pp. 79 – 87, 2002.

[55] M. Spear, "The influence of contrast effects upon teachers' marks," *Educational Research*, vol. 39, pp. 229–233, 1997.

[56] UK Public General Acts, "Education act 1918," 1918.

[57] UK Parliament, "Education act 1944 (7 & 8 geo.6 c.31)," https://www.legislation.gov.uk/ukpga/Geo6/7-8/31/contents, 1944, accessed: 2025-09-18.

[58] BBC News. (2004) Primary school tests toned down. [Online]. Available: http://news.bbc.co.uk/1/hi/education/3656244.stm

[59] BBC. (2008) Tests scrapped for 14-year-olds. [Online]. Available: http://news.bbc.co.uk/1/hi/education/7669254.stm

[60] Department for Education. (2013) Assessing without levels. [Online]. Available: https://webarchive.nationalarchives.gov.uk/ukgwa/20130802141012/https://www.education.gov.uk/schools/teachingandlearning/curriculum/nationalcurriculum2014/a00225864/assessing-without-levels

[61] H. Rugg, "Teachers' marks and the reconstruction of the marking system," *The Elementary School Journal*, vol. 18, pp. 701 – 719, 1918.

[62] J. Cain, M. S. Medina, F. Romanelli, and A. Persky, "Deficiencies of traditional grading systems and recommendations for the future," *American Journal of Pharmaceutical Education*, vol. 86, 2021.

[63] V. Crisp, "Exploring the nature of examiner thinking during the process of examination marking," *Cambridge Journal of Education*, vol. 38, pp. 247 – 264, 2008.

[64] A. P. de Moira, C. Massey, J. Baird, and M. Morrissy, "Marking consistency over time," *Research in Education*, vol. 67, pp. 79 – 87, 2002.

[65] A. Scharaschkin and J. Baird, "The effects of consistency of performance on a-level examiners' judgements of standards," *British Educational Research Journal*, vol. 26, pp. 343–357, 2000.

[66] Z. Primack, J. Splett, and J. Graham, "Teacher stress, teacher unintentional bias, and teacher well-being before and during covid-19," *UF Journal of Undergraduate Research*, 2023.

[67] G. Zanga and E. D. Gioannis, "Discrimination in grading: A scoping review of studies on teachers' discrimination in school," *Studies in Educational Evaluation*, 2023.

[68] K. Willey and A. Gardner, "Improving the standard and consistency of multi-tutor grading in large classes," 2010.

[69] T. Guskey, "Addressing inconsistencies in grading practices," *Phi Delta Kappan*, vol. 105, pp. 52 – 57, 2024.

[70] E. Pitt and N. Winstone, "The impact of anonymous marking on students' perceptions of fairness, feedback and relationships with lecturers," *Assessment & Evaluation in Higher Education*, vol. 43, no. 7, pp. 1183–1193, 2018.

[71] Department for Education, "Teacher workload survey 2019: Main report," p. 10, 2019, accessed: 2024-12-28. [Online]. Available: https://assets.publishing.service.gov.uk/media/5e12fcb7e5274a0f9e82e4fd/teacher_workload_survey_2019_main_report_amended.pdf

[72] National Education Union, "Workload advice," 2024, accessed: 2024-12-28. [Online]. Available: https://neu.org.uk/advice/your-rights-work/contracts-and-working-hours/workload-and-working-time/workload-advice

[73] Department for Education, "School workload reduction toolkit," 2022, retrieved December 29, 2024. [Online]. Available: https://www.gov.uk/guidance/school-workload-reduction-toolkit#how-to-use-the-toolkit

[74] Department for Education (DfE), "Working lives of teachers and leaders: Wave 3 summary report," 2024, accessed: 2024-12-28. [Online]. Available: https://assets.publishing.service.gov.uk/media/674ddf916f6baefc2a9ca1aa/Working_lives_of_teachers_and_leaders_wave_3_summary_report.pdf

[75] J. Jerrim and S. Sims, "When is high workload bad for teacher wellbeing? accounting for the non-linear contribution of specific teaching tasks," *Teaching and Teacher Education*, vol. 105, p. 103395, 2021.

[76] S. Erturk, W. A. van Tilburg, and E. R. Igou, "Off the mark: Repetitive marking undermines essay evaluations due to boredom," *Motivation and Emotion*, vol. 46, no. 2, pp. 264–275, 2022.

[77] C. Senanayake and D. Asanka, "Rubric based automated short answer scoring using large language models (llms)," *2024 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, vol. 7, pp. 1–6, 2024.

[78] S. Bloxham, P. Boyd, and S. Orr, "Mark my words: the role of assessment criteria in uk higher education grading practices," *Studies in Higher Education*, vol. 36, pp. 655 – 670, 2011.

[79] H. Hausdorff and S. Farr, "The effect of grading practices on the marks of gifted sixth grade children," *Journal of Educational Research*, vol. 59, pp. 169–172, 1965.

[80] L. Norton, S. Floyd, and B. Norton, "Lecturers' views of assessment design, marking and feedback in higher education: a case for professionalisation?" *Assessment & Evaluation in Higher Education*, vol. 44, pp. 1209–1221, 2019.

[81] R. Raaper, "Academic perceptions of higher education assessment processes in neoliberal academia," *Critical Studies in Education*, vol. 57, no. 2, pp. 175–190, 2016.

[82] J. R. Spencer and D. Horn, "The three most important words in faculty workload: Transparency, transparency, transparency," *The Department Chair*, vol. 34, no. 1, pp. 1–3, 2023.

[83] F. Ilahi, T. Manzoor, and I. Elahi, "Enhancing assessment integrity: A critical analysis of transparency and fairness in marking process at university of sargodha," *Journal of Education and Social Studies*, 2024.

[84] R. Watermeyer, R. Bolden, C. Knight, and T. Crick, "Academic anomie: implications of the 'great resignation' for leadership in post-COVID higher education," *Higher Education*, vol. 89, pp. 1215–1233, 2025.

[85] A. Hasan and B. Jones, "Assessing the assessors: investigating the process of marking essays," *Frontiers in Oral Health*, vol. 5, 2024.

[86] H. Torrance and J. Pryor, *Investigating formative assessment: Teaching, learning and assessment in the classroom*. McGraw-Hill Education (UK), 1998.

[87] P. Black and C. Harrison, "Feedback in questioning and marking: The science teacher's role in formative assessment," *School science review*, vol. 82, no. 301, pp. 55–61, 2001.

[88] OECD. (2005) Formative assessment: Improving learning in secondary classrooms. [Online]. Available: https://www.oecd.org/education/ceri/35661078.pdf

[89] D. William, "National curriculum assessment arrangements," *British Journal for Curriculum and Assessment*, vol. 1, pp. 8–12, 1990.

[90] TES Editorial, "Directed improvement and reflection time (dirt) and teacher feedback," 2023, accessed: 2024-10-03. [Online]. Available: https://www.tes.com/magazine/teaching-learning/general/directed-improvement-and-reflection-time-dirt-teacher-feedback

[91] Joint Council for Qualifications, "Suspected malpractice: Policies and procedures 2024–2025," 2024. [Online]. Available: https://www.jcq.org.uk/wp-content/uploads/2024/08/Malpractice_Sep24_FINAL.pdf

[92] Ofqual, "Malpractice in GCSE, AS and A level: summer 2024 exam series," 2024, accessed: 2025-09-18. [Online]. Available: https://www.gov.uk/government/statistics/malpractice-in-gcse-as-and-a-level-summer-2024-exam-series/malpractice-in-gcse-as-and-a-level-summer-2024-exam-series

[93] J. M. Olson and R. Krysiak, "Rubrics as tools for effective assessment of student learning and program quality," in *Curriculum Development and Online Instruction for the 21st Century*. IGI Global, 2021, pp. 173–200.

[94] T. Sarı, F. Nayır, and Şenel Poyrazlı, "Educare, educere, or holistic? exploring researchers' educational approaches in education for sustainable development research," *Sustainable Development*, 2025.

[95] G. Cox, J. Morrison, and B. Brathwaite, "The rubric: an assessment tool to guide students and markers," in *1ST INTERNATIONAL CONFERENCE ON HIGHER EDUCATION ADVANCES (HEAD'15)*. Editorial Universitat Politècnica de València, 2015, pp. 26–32.

[96] K. Sambell, S. Brown, and P. Race, "Assessment to support student learning: eight challenges for 21st century practice," *All Ireland Journal of Teaching and Learning in Higher Education (AISHE-J) Creative Commons Attribution-NonCommercial-ShareAlike*, vol. 11, no. 2, 2019.

[97] A. Jonsson and G. Svingby, "The use of scoring rubrics: Reliability, validity, and educational consequences," *Educational Research Review*, vol. 2, pp. 130–144, 2007.

[98] S. Brookhart and F. Chen, "The quality and effectiveness of descriptive rubrics," *Educational Review*, vol. 67, pp. 343 – 368, 2015.

[99] S. Yune, S. Y. Lee, S. Im, B. Kam, and S. Baek, "Holistic rubric vs. analytic rubric for measuring clinical performance levels in medical students," *BMC Medical Education*, vol. 18, 2018.

[100] G. Badia, "Holistic or analytic rubrics? grading information literacy instruction," *College & Undergraduate Libraries*, vol. 26, pp. 109 – 116, 2019.

[101] C. Tomás, E. Whitt, R. Lavelle-Hill, and K. E. Severn, "Modeling holistic marks with analytic rubrics," *Frontiers in Education*, 2019.

[102] J. Chen, H. Yang, and C. Han, "Holistic versus analytic scoring of spoken-language interpreting: a multi-perspectival comparative analysis," *The Interpreter and Translator Trainer*, vol. 16, pp. 558 – 576, 2022.

[103] A. Cockett and C. Jackson, "The use of assessment rubrics to enhance feedback in higher education: An integrative literature review." *Nurse education today*, vol. 69, pp. 8–13, 2018.

[104] Y. M. Reddy and H. Andrade, "A review of rubric use in higher education," *Assessment & Evaluation in Higher Education*, vol. 35, pp. 435 – 448, 2010.

[105] E. Panadero and A. Jonsson, "A critical review of the arguments against the use of rubrics," *Educational Research Review*, vol. 30, p. 100329, 2020.

[106] R. Smit, P. Bachmann, V. Blum, T. Birri, and K. Hess, "Effects of a rubric for mathematical reasoning on teaching and learning in primary school," *Instructional Science*, vol. 45, pp. 603–622, 2017.

[107] Z. Chan and S. S. M. Ho, "Good and bad practices in rubrics: the perspectives of students and educators," *Assessment & Evaluation in Higher Education*, vol. 44, pp. 533 – 545, 2019.

[108] K. Martens, "Rubrics in program evaluation," *Evaluation Journal of Australasia*, vol. 18, pp. 21 – 44, 2018.

[109] G. F. Shaw, "Introducing rubrics to physical education teacher candidates," *Journal of Physical Education, Recreation and Dance*, vol. 85, pp. 31 – 37, 2014.

[110] C. Hack, "Analytical rubrics in higher education: A repository of empirical data," *British Journal of Educational Technology*, vol. 46, no. 5, pp. 924–927, 2015.

[111] I. Nisbet and S. D. Shaw, *Is Assessment Fair?* SAGE Publications Ltd, 2020.

[112] C. Knight, C. Conn, T. Crick, and S. Brooks, "Divergences in the Framing of Inclusive Education across the UK: A Four Nations Critical Policy Analysis," *Educational Review*, vol. 77, no. 2, pp. 495–511, 2025.

[113] T. Crick, "Covid-19 and digital education: A catalyst for change?" *Itnow*, vol. 63, no. 1, pp. 16–17, 2021.

[114] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. Slade, A. Jeyaraj, A. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, M. Ahuja, M. Al-Bashrawi, A. S. Al-Busaidi, J. Balakrishnan, Y. Barlette, S. Basu, I. Bose, L. Brooks, D. Buhalis, L. Carter, S. Chowdhury, T. Crick, S. W. Cunningham, G. H. Davies, R. M. Davison, R. Dé, D. Dennehy, Y. Duan, R. Dubey, V. Dutot, R. Dwivedi, J. S. Edwards, C. Flavián, R. Gauld, V. Grover, M.-C. Hu, M. Janssen, P. Jones, I. Junglas, S. Khorana, S. Kraus, K. R. Larsen, P. Latreille, S. Laumer, F. T. Malik, A. Mardani, M. Mariani, S. Mithas, E. Mogaji, J. Nord, S. O'Connor, F. Okumus, M. Pagani, N. Pandey, I. O. Pappas, J. Pries-Heje, S. Papagiannidis, N. Pathak, R. Raman, N. P. Rana, S.-V. Rehm, S. Ribeiro-Navarrete, A. Richter, F. Rowe, S. Sarker, B. Stahl, M. Tiwari, W. van der Aalst, V. Venkatesh, G. Viglia, M. Wade, P. Walton, J. Wirtz, and R. Wright, ""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*, vol. 71, no. 102642, 2023.

[115] Z. Swiecki, H. Khosravi, G. Chen, R. Martinez-Maldonado, J. M. Lodge, S. Milligan, N. Selwyn, and D. Gašević, "Assessment in the age of artificial intelligence," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100075, 2022.

[116] R. Morris, S. Gorard, B. See, and N. Siddiqui, "Can a code-based approach to marking and feedback reduce teachers' workload? an evaluation of the flash marking intervention," *Oxford Review of Education*, vol. 50, pp. 552 – 569, 2023.

[117] A. Rasooli, H. Zandi, and C. DeLuca, "Re-conceptualizing classroom assessment fairness: A systematic meta-ethnography of assessment literature and beyond," *Studies in Educational Evaluation*, vol. 56, pp. 164–181, 2018.

[118] R. Tierney, "Fairness as a multifaceted quality in classroom assessment," *Studies in Educational Evaluation*, vol. 43, pp. 55–69, 2014.

[119] M. Bamber, "The impact on stakeholder confidence of increased transparency in the examination assessment process," *Assessment & Evaluation in Higher Education*, vol. 40, pp. 471 – 487, 2015.

[120] A. Rasooli, C. DeLuca, A. Rasegh, and S. Fathi, "Students' critical incidents of fairness in classroom assessment: an empirical study," *Social Psychology of Education*, pp. 1–22, 2019.

[121] A. Pollitt, "Comparative judgment for assessment," *International Journal of Technology and Design Education*, vol. 22, no. 2, pp. 157–170, 2012.

[122] R. Song, Q. Guo, R. Zhang, G. Xin, J.-R. Wen, Y. Yu, and H. Hon, "Select-the-best-ones: A new way to judge relative relevance," *Inf. Process. Manag.*, vol. 47, pp. 37–52, 2011.

[123] L. Magowan, "Centre assessment grades in 2020: A natural experiment for investigating bias in teacher judgements," *Journal of Computational Social Science*, vol. 6, no. 2, pp. 609–653, 2023.

[124] C. Gonsalves and Z. Lin, "Clear in advance to whom? exploring 'transparency'of assessment practices in uk higher education institution assessment policy," *Studies in Higher Education*, vol. 50, no. 7, pp. 1454–1470, 2025.

[125] S. Norton and K. Hack, "Framework for Enhancing Assessment in Higher Education," Advance HE, Tech. Rep., January 2024.

[126] Quality Assurance Agency for Higher Education, "Uk quality code for higher education: Advice and guidance on assessment," 2018, accessed: 2025-03-08. [Online]. Available: https://www.qaa.ac.uk/the-quality-code/advice-and-guidance/assessment

[127] Office for Students, "The regulatory framework for higher education in england," 2022, accessed: 2025-03-08. [Online]. Available: https://www.officeforstudents.org.uk/publications/the-regulatory-framework-for-higher-education-in-england/

[128] S. Walker, "Trends in assessment in higher education: considerations for policy and practice," Jisc, Tech. Rep., January 2025.

[129] M. Ragolane, S. Patel, and P. Salikram, "Ai versus human graders: Assessing the role of large language models in higher education," *Asian Journal of Education and Social Studies*, 2024.

[130] S. Bloxham, "Marking and moderation in the UK: false assumptions and wasted resources," *Assessment & Evaluation in Higher Education*, vol. 34, pp. 209–220, 2009.

[131] A. Pollitt, "Comparative judgement for assessment," *International Journal of Technology and Design Education*, vol. 22, no. 2, pp. 157–170, 2012.

[132] T. Benton and T. Gallagher, "Is comparative judgement just a quick form of multiple marking," *Research Matters: A Cambridge Assessment Publication (26)*, pp. 22–28, 2018.

[133] A. Pollitt and N. L. Murray, "What raters really pay attention to," *Studies in Language Testing*, vol. 3, pp. 74–91, 1996.

[134] D. R. Hunter, "Mm algorithms for generalized bradley-terry models," *The annals of statistics*, vol. 32, no. 1, pp. 384–406, 2004.

[135] L. L. Thurstone, "A law of comparative judgment." *Psychological review*, vol. 34, no. 4, p. 273, 1927.

[136] T. Coenen, L. Coertjens, P. Vlerick, M. Lesterhuis, A. V. Mortier, V. Donche, P. Ballon, and S. De Maeyer, "An information system design theory for the comparative judgement of competences," *European Journal of Information Systems*, vol. 27, no. 2, pp. 248–261, 2018.

[137] J. Arbuckle and J. H. Nugent, "A general procedure for parameter estimation for the law of comparative judgement," *British Journal of Mathematical and Statistical Psychology*, vol. 26, no. 2, pp. 240–260, 1973.

[138] R. M. Furr, *Psychometrics: an introduction*. SAGE publications, 2021.

[139] G. A. Gescheider, *Psychophysics: the fundamentals*. Psychology Press, 2013.

[140] A. Pollitt, "Let's stop marking exams," in *IAEA Conference*, 2004, University of Cambridge Local Examinations Syndicate.

[141] S. R. Bartholomew, G. J. Strimel, and E. Yoshikawa, "Using adaptive comparative judgment for student formative feedback and learning during a middle school design project," *International Journal of Technology and Design Education*, vol. 29, no. 2, pp. 363–385, 2019.

[142] D. Kahneman and A. Tversky, "Prospect Theory: An Analysis of Decision Under Risk," in *Handbook of the Fundamentals of Financial Decision Making: Part I*. World Scientific, 2013, pp. 99–127.

[143] D. R. Sadler, "Formative assessment and the design of instructional systems," *Instructional Science*, vol. 18, pp. 119–144, 1989.

[144] T. Bramley, "Paired comparison methods," in *Techniques for monitoring the comparability of examination standards*, 2007, pp. 246–300.

[145] O. Chen, F. Paas, and J. Sweller, "A Cognitive Load Theory Approach to Defining and Measuring Task Complexity Through Element Interactivity," *Educational Psychology Review*, vol. 35, no. 63, 2023.

[146] D. Andrich, "A rating formulation for ordered response categories," *Psychometrika*, vol. 43, no. 4, pp. 561–573, 1978.

[147] S. Holmes, B. Black, and C. Morin, "Marking reliability studies 2017: Rank ordering versus marking – which is more reliable?" Ofqual, Tech. Rep., January 2020.

[148] T. Bramley, "Investigating the reliability of adaptive comparative judgment," *Cambridge Assessment, Cambridge*, vol. 36, 2015.

[149] A. Pinot de Moira, C. Wheadon, and D. Christodoulou, "The classification accuracy and consistency of comparative judgement of writing compared to rubric-based teacher assessment," *Research in Education*, vol. 113, no. 1, pp. 25–40, 2022.

[150] S. Verhavert, S. De Maeyer, V. Donche, and L. Coertjens, "Scale Separation Reliability: What Does It Mean in the Context of Comparative Judgment?" *Applied Psychological Measurement*, vol. 42, no. 6, pp. 428–445, 2018.

[151] J. T. Steedle and S. Ferrara, "Evaluating Comparative Judgment as an Approach to Essay Scoring," *Applied Measurement in Education*, vol. 29, no. 3, pp. 211–223, 2016.

[152] D. E. Hinkle, W. Wiersma, and S. G. Jurs, *Applied Statistics for the Behavioural Sciences*, 6th ed. Houghton Mifflin, 2002.

[153] C. Wheadon, P. Barmby, D. Christodoulou, and B. Henderson, "A comparative judgement approach to the large-scale assessment of primary writing in england," *Assessment in Education: Principles, Policy & Practice*, vol. 27, no. 1, pp. 46–64, 2020.

[154] I. Jones and B. Davies, "Comparative judgement in education research," *International Journal of Research & Method in Education*, 2022.

[155] A. Pollitt, "The method of adaptive comparative judgement," *Assessment in Education: Principles, Policy & Practice*, vol. 19, no. 3, pp. 281–300, 2012.

[156] T. Bramley and S. Vitello, "The effect of adaptivity on the reliability coefficient in adaptive comparative judgement," *Assessment in Education: Principles, Policy & Practice*, vol. 26, no. 1, pp. 43–58, 2019.

[157] N. Seery, D. Canty, and P. Phelan, "The validity and value of peer assessment using adaptive comparative judgement in design driven practical education," *International Journal of Technology and Design Education*, vol. 22, no. 2, pp. 205–226, 2012.

[158] S. Bartholomew and M. D. Jones, "A systematized review of research with adaptive comparative judgment (ACJ) in higher education," *International Journal of Technology and Design Education*, vol. 32, pp. 1159–1190, 2021.

[159] J. Buckley, N. Seery, and R. Kimbell, "A review of the valid methodological use of adaptive comparative judgment in technology education research," vol. 7, 2022.

[160] N. Mentzer, W. Lee, and S. Bartholomew, "Examining the validity of adaptive comparative judgment for peer evaluation in a design thinking course," vol. 6, 2021.

[161] M.-J. Bisson, C. Gilmore, M. Inglis, and I. Jones, "Measuring conceptual understanding using comparative judgement," *International Journal of Research in Undergraduate Mathematics Education*, vol. 2, no. 2, pp. 141–164, 2016.

[162] N. Marshall, K. Shaw, J. Hunter, and I. Jones, "Assessment by comparative judgement: An application to secondary statistics and english in new zealand," *New Zealand Journal of Educational Studies*, vol. 55, no. 1, pp. 49–71, 2020.

[163] A. Gray, A. A. Rahat, T. Crick, S. Lindsay, and D. Wallace, "Using Elo rating as a metric for comparative judgement in educational assessment," in *Proceedings of 6th International Conference on Education and Multimedia Technology (ICEMT 2022)*, 2022, pp. 272–278.

[164] A. Agresti, *Categorical data analysis*. John Wiley & Sons, 2003.

[165] R. Baker and P. Scarf, "Modifying bradley–terry and other ranking models to allow ties," *IMA Journal of Management Mathematics*, vol. 32, no. 4, pp. 451–463, 2021.

[166] O. A. Martin, R. Kumar, and J. Lao, *Bayesian Modeling and Computation in Python*, 2021.

[167] A. B. Downey, *Think Bayes*. O'Reilly Media, 2021.

[168] D. Sivia and J. Skilling, *Data analysis: a Bayesian tutorial*. Oxford University Press, 2006.

[169] K. Tsukida and M. R. Gupta, "How to analyze paired comparison data," Department of Electrical Engineering, University of Washington, Tech. Rep. UWEETR-2011-0004, 2011.

[170] J. Wainer, "A Bayesian Bradley-Terry model to compare multiple ML algorithms on multiple data sets," *arXiv*, 2022.

[171] S. De Maeyer, "Bayesian analysis of comparative judgement data," https://svendemaeyer.netlify.app/posts/2021-01-18-bayesian-analysis-of-comparative-judgement-data/, January 2021.

[172] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[173] J. R. Quinlan, "Induction of decision trees," in *Machine learning*. Springer, 1986, pp. 81–106.

[174] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Harley, T. Lillicrap, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," *International conference on machine learning (ICML)*, 2016.

[175] C. M. Bishop, "Pattern recognition and machine learning: Springer science+ business media," *LLC, NY, USA*, 2006.

[176] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2006.

[177] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems (NeurIPS)*, vol. 30, 2017.

[178] K. G. Jamieson and R. Nowak, "Active ranking using pairwise comparisons," *Advances in neural information processing systems*, vol. 24, 2011.

[179] R. Heckel, N. B. Shah, K. Ramchandran, and M. J. Wainwright, "Active ranking from pairwise comparisons and when parametric assumptions do not help," 2019.

[180] A. Tversky and J. E. Russo, "Substitutability and similarity in binary choices," *Journal of Mathematical psychology*, vol. 6, no. 1, pp. 1–12, 1969.

[181] Z. Wang and Y. Feng, "Harmonizing form and function: The evolution, principles, and future of interactive design," *Applied and Computational Engineering*, 2024.

[182] A. Dix, "What is human-computer interaction (hci)," *Interaction Design Foundation. https://acortar. link/mETsan*, 2016.

[183] H. Hasyim and M. Bakri, "Advancements in human-computer interaction: A review of recent research," *Advances: Jurnal Ekonomi & Bisnis*, 2024.

[184] A. L. Kotian, R. Nandipi, U. M, U. S, Varshauk, and V. G. T, "A systematic review on human and computer interaction," *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pp. 1214–1218, 2024.

[185] B. Shneiderman, *Human-Centered AI*. Oxford University Press, 2022.

[186] C. R. Becker, *Learn Human-Computer Interaction: Solve human problems and focus on rapid prototyping and validating solutions through user testing*. Packt Publishing Ltd, 2020.

[187] A. Blandford, "Eliciting people's conceptual models of activities and systems," *International Journal of Conceptual Structures and Smart Applications (IJCSSA)*, vol. 1, no. 1, pp. 1–17, 2013.

[188] A. Blandford, T. R. Green, D. Furniss, and S. Makri, "Evaluating system utility and conceptual fit using cassm," *International Journal of Human-Computer Studies*, vol. 66, no. 6, pp. 393–409, 2008.

[189] A. Blandford and G. Rugg, "A case study on integrating contextual information with analytical usability evaluation," *International journal of human-computer studies*, vol. 57, no. 1, pp. 75–99, 2002.

[190] G. B. W. Sharrock, "Studies of work and the workplace in hci: Concepts and techniques," 2009.

[191] P. Dourish, "Implications for design," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006, pp. 541–550.

[192] A. Kamsin, A. Blandford, and A. L. Cox, "Personal task management: my tools fall apart when i'm very busy!" in *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, 2012, pp. 1369–1374.

[193] A. L. Strauss, *The discovery of grounded theory: Strategies for qualitative research*. Routledge, 2017.

[194] D. Furniss, A. Blandford, and P. Curzon, "Confessions from a grounded theory phd: experiences and lessons learnt," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 113–122.

[195] H. Yoo, J. Han, S.-Y. Ahn, and A. Oh, "Dress: Dataset for rubric-based essay scoring on efl writing," *arXiv preprint arXiv:2402.16733*, 2024.

[196] J. Neyman, "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection," in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 123–150.

[197] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, p. 81–93, 1938.

[198] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing Top k Lists," *SIAM Journal on Discrete Mathematics*, vol. 17, no. 1, pp. 134–160, 2003.

[199] S. Kullback, "Kullback-leibler divergence," *Tech. Rep.*, 1951.

[200] M. L. Menéndez, J. A. Pardo, L. Pardo, and M. d. C. Pardo, "The jensen-shannon divergence," *Journal of the Franklin Institute*, vol. 334, no. 2, pp. 307–318, 1997.

[201] T. W. MacFarland, J. M. Yates *et al.*, *Introduction to nonparametric statistics for the biological sciences using R*. Springer, 2016.

[202] O. J. Dunn, "Multiple comparisons among means," *Journal of the American statistical association*, vol. 56, no. 293, pp. 52–64, 1961.

[203] E. J. Hughes, "Evolutionary Multi-objective Ranking with Uncertainty and Noise," in *International Conference on Evolutionary Multi-Criterion Optimization (EMO 2001)*, ser. LNCS, 2001, vol. 1993, pp. 329–343.

[204] L. C. Andrews, *Special Functions of Mathematics for Engineers*. Oxford University Press, 1998.

[205] D. Bertsekas and J. N. Tsitsiklis, *Introduction to probability*. Athena Scientific, 2008, vol. 1.

[206] R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen *et al.*, "Bayesian statistics and modelling," *Nature Reviews Methods Primers*, vol. 1, no. 1, 2021.

[207] R. McElreath, "Statistical Rethinking: A Bayesian Course with Examples in R and STAN," 2020.

[208] B. Lambert, "A student's guide to bayesian statistics," *A Student's Guide to Bayesian Statistics*, pp. 1–520, 2018.

[209] J. N. Pritikin, "An exploratory factor model for ordinal paired comparison indicators," *Heliyon*, vol. 6, no. 9, p. e04821, 2020.

[210] D. Fink, "A Compendium of Conjugate Priors," Tech. Rep., May 1997.

[211] D. J. C. Mackay, "Introduction to Monte Carlo methods," in *Learning in Graphical Models*. Springer, 1998, pp. 175–204.

[212] E. Koehler, E. Brown, and S. J.-P. Haneuse, "On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses," *The American Statistician*, vol. 63, no. 2, pp. 155–162, 2009.

[213] P. Thiagarajan and S. Ghosh, "Jensen-Shannon Divergence Based Novel Loss Functions for Bayesian Neural Networks," *arXiv*, 2023.

[214] B. Settles, "Active Learning Literature Survey," University of Wisconsin-Madison, Tech. Rep. Computer Sciences Technical Report 1648, January 2010.

[215] B. P. Knijnenburg and M. C. Willemsen, "Evaluating Recommender Systems with User Experiments," in *Recommender Systems Handbook*. Springer, 2015, pp. 309–352.

[216] S. Das, W.-K. Wong, T. Dietterich, A. Fern, and A. Emmott, "Incorporating Expert Feedback into Active Anomaly Discovery," in *IEEE 16th International Conference on Data Mining (ICDM 2016)*, 2016, pp. 853–858.

[217] D. J. C. MacKay, "Information-Based Objective Functions for Active Data Selection," *Neural Computation*, vol. 4, no. 4, pp. 590–604, 1992.

[218] X. Zhan, H. Liu, Q. Li, and A. B. Chan, "A comparative survey: Benchmarking for pool-based active learning," in *Proceedings of 30th International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021, pp. 4679–4686.

[219] D. D. Lewis, "A sequential algorithm for training text classifiers: Corrigendum and additional data," *ACM SIGIR Forum*, vol. 29, no. 2, pp. 13–19, 1995.

[220] A. V. Lazo and P. Rathie, "On the entropy of continuous probability distributions (Corresp.)," *IEEE Transactions on Information Theory*, vol. 24, no. 1, pp. 120–122, 1978.

[221] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[222] R. G. Miller, "Simultaneous statistical inference," 1981.

[223] A. Gray, A. Rahat, T. Crick, and S. Lindsay, "A bayesian active learning approach to comparative judgement within education assessment," *Computers and Education: Artificial Intelligence*, p. 100245, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666920X24000481

[224] R. Watermeyer, L. Phipps, D. Lanclos, and C. Knight, "Generative AI and the Automating of Academia," *Postdigital Science and Education*, 2023.

[225] D. Laming, "The relativity of 'absolute'judgements," *British Journal of Mathematical and Statistical Psychology*, vol. 37, no. 2, pp. 152–183, 1984.

[226] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: The method of paired comparisons," *Biometrika*, vol. 39, no. 3-4, pp. 324–345, 1952.

[227] R. H. Ashton, "A review and analysis of research on the test–retest reliability of professional judgment," *Journal of Behavioral Decision Making*, vol. 13, no. 3, pp. 277–294, 2000.

[228] G. Kinnear, I. Jones, and B. Davies, "Comparative judgement as a research tool: a meta-analysis of application and reliability," Center for Open Science, Tech. Rep., 2025.

[229] Q. Yu and K. M. Quinn, "A multidimensional pairwise comparison model for heterogeneous perceptions with an application to modelling the perceived truthfulness of public statements on covid-19," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 185, no. 3, pp. 1049–1073, 2022.

[230] R. A. Hefner Jr, *Extensions of the law of comparative judgment to discriminable and multidimensional stimuli*. University of Michigan, 1959.

[231] C. Sangwin and G. Kinnear, "Investigating insight and rigour as separate constructs in mathematical proof," *Research in Mathematics Education*, pp. 1–29, 2024.

[232] R. Sickinger, T. Brunfaut, and J. Pill, "Comparative judgement for evaluating young learners' efl writing performances: Reliability and teacher perceptions of holistic and dimension-based judgements," *Language Testing*, vol. 42, no. 2, pp. 137–166, 2025.

[233] B. G. Lindsay, "Mixture models: theory, geometry, and applications." Ims, 1995.

[234] M. Abramowitz, "I. a. stegun (editors), handbook of mathematical functions," *Applied Mathematics Series*, 1972.

[235] G. A. Korn and T. M. Korn, *Mathematical handbook for scientists and engineers: definitions, theorems, and formulas for reference and review*. Courier Corporation, 2000.

[236] W. J. Morokoff and R. E. Caflisch, "Quasi-monte carlo integration," *Journal of computational physics*, vol. 122, no. 2, pp. 218–230, 1995.

[237] N. A. Smith and R. W. Tromble, "Sampling uniformly from the unit simplex," *Johns Hopkins University, Tech. Rep*, vol. 29, 2004.

[238] L.-C. Velasco-Martinez, J.-C. Tojar-Hurtado *et al.*, "Transparency in evaluation through the use of rubrics in university subjects," 2019.

[239] I. Jones and B. Davies, "Comparative judgement in education research," *International Journal of Research & Method in Education*, vol. 47, no. 2, pp. 170–181, 2024.

[240] A. Gray, A. Rahat, T. Crick, and S. Lindsay, "Bayesian active learning for comparative judgement: Estimating reliability and managing multiple criteria with applications in educational assessment," 2025. [Online]. Available: https://arxiv.org/abs/2503.00479

[241] V. Crisp, "The judgement processes involved in the moderation of teacher-assessed projects," *Oxford Review of Education*, vol. 43, no. 1, pp. 19–37, 2017.

[242] S. Bloxham, C. Hughes, and L. A. and, "What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices," *Assessment & Evaluation in Higher Education*, vol. 41, no. 4, pp. 638–653, 2016.

[243] I. Jones, M.-J. Bisson, C. Gilmore, and M. Inglis, "Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help?" *British Educational Research Journal*, 2019.

[244] A. Gray, A. A. Rahat, T. Crick, S. Lindsay, and D. Wallace, "Using elo rating as a metric for comparative judgement in educational assessment," in *Proceedings of the 6th International Conference on Education and Multimedia Technology*, ser. ICEMT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 272–278. [Online]. Available: https://doi.org/10.1145/3551708.3556204

[245] C. Attig, N. Rauh, T. Franke, and J. F. Krems, "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)*, 2017, pp. 3–14.

[246] R. E. Caflisch, "Monte carlo and quasi-monte carlo methods," *Acta Numerica*, vol. 7, pp. 1–49, 1998.

Appendix A

HCI Questionnaire

Introduction

We are conducting this survey to understand your experiences with three different marking methods: Traditional Marking, Bayesian Comparative Judgement, and Bayesian Comparative Judgement. Your feedback is valuable and will help us improve these methods. This survey should take about 10-15 minutes to complete.

Experience with Marking Methods

For each of the following sections, please rate your experience on a scale of 1 (Very Poor) to 5 (Excellent).

Definitions:

- Transparent: The clarity and openness of the processes, criteria, and decisions involved in evaluating students' work. It ensures that students, teachers, and other stakeholders understand how judgements are made, promoting fairness and trust in the outcomes.

- Traditional absolute marking: Evaluating students' work against a fixed set of criteria or a predetermined mark scheme, where grades are awarded based on the extent to which these criteria are met.

- Bayesian Comparative Judgement (BCJ): Assessing students' work by comparing pairs of pieces and using a probabilistic model to estimate the quality of each, based on the collective judgements made.
- Multi-criteria BCJ: Multi-criteria comparative judgement involves comparing pairs of students' work across multiple criteria or dimensions, allowing for a holistic evaluation that simultaneously considers various aspects of quality.

Traditional Marking

1. How would you rate the ease of use of Traditional Marking (1-5) and why?
2. How transparent do you find the process of Traditional Marking (1-5) and why?
3. How accurate do you find the results of Traditional Marking?
4. What were your initial impressions of the marking approach?
5. What do you think is happening when you use traditional marking methods?
6. Do you enjoy using this approach to mark the student's work?
7. What alterations would you want to see to it?
8. Does this approach to marking work fit with how you think about student work?

Bayesian Comparative Judgement

9. How would you rate the ease of use of Bayesian Comparative Judgement (1-5) and why?
10. How transparent do you find the process of Bayesian Comparative Judgement (1-5) and why?
11. How confident are you in the ranks derived from the Bayesian Comparative Judgement (1-5) and why?
12. What were your initial impressions of the marking approach?
13. What do you think is happening when you use Bayesian Comparative Judgement?

14. Would you recommend this approach over traditional marking? Why?
15. What alterations would you want to see to it?
16. Does this approach to marking work fit with how you think about student work?

Multi-Criteria Bayesian Comparative Judgement

17. How would you rate the ease of use of the multi-criteria Bayesian Comparative Judgement (1-5) and why?
18. How transparent do you find the process of the multi-criteria Bayesian Comparative Judgement (1-5) and why?
19. How confident are you in the ranks derived by using the multi-criteria Bayesian Comparative Judgement (1-5) and why?
20. What were your initial impressions of the marking approach?
21. What do you think is happening when you use multi-criteria Bayesian Comparative Judgement?
22. Would you recommend this approach over traditional marking? Why?
23. What alterations would you want to see to it?
24. Does this approach to marking work fit with how you think about student work?

Additional Feedback

25. Which marking method did you prefer and why?
26. Which approach do you have the most confidence in their rankings of the work and Why?

Thank you for your time and feedback!

Appendix B

Workshop

1. Welcome and Introduction

- Recap the three marking methods: Traditional Marking, Bayesian Comparative Judgement (BCJ), and Multi-criteria Bayesian Comparative Judgement (MDBCJ).
- Outline the objectives of the workshop:
  - Explore participants' experiences with the marking methods.
  - Present marking outcomes, including metrics such as tau scores and rank comparisons.
  - Discuss the implications of these outcomes on trust and transparency.
  - Evaluate transparency and trustworthiness from a student perspective.

2. Assumptions and Processes

- Facilitate an exploration of participants' assumptions about each method:
  - What assumptions did you make about the marking processes?
  - Did the results or tau scores challenge these assumptions?
  - How do the results reflect fairness or accuracy in marking?
- Draw connections between assumptions, outcomes, and trust in the methods.

3. Group Reflection: Experiences with Marking Methods

- Divide participants into small groups to discuss:
  - How does your experience with the marking methods align with your performance outcomes?
  - Which method do you feel most confident using, and why?
  - Did the performance comparisons (tau scores) change your perceptions of the methods?
- Summarise group discussions and share key insights.

4. Presentation of Marking Outcomes

- Present participants' marking outcomes:
  - **Performance Metrics:** Show how participants' rankings align with the target rank using the tau score.
  - Compare individual ranks against each other.
- Facilitate discussion:
  - What do these metrics reveal about the reliability of each method?
  - How do the results align with your expectations?
  - Does seeing the comparative performance affect your trust in the methods?

5. Trustworthiness and Transparency Discussion (30 minutes)

- Facilitate an open discussion using prompts:
  - Which method do you trust the most after seeing the results? Why?
  - From a student perspective:
    - Which method best conveys fairness and transparency?
    - Would knowing about tau scores or ranking comparisons build or undermine student trust?

6. Reflection and Actionable Feedback

- Ask participants to reflect individually on:
  - Key insights about the strengths and weaknesses of each method.
  - Suggestions to improve usability, trust, or transparency.
  - How seeing their performance outcomes impacts their overall views.
- Collect feedback using sticky notes or a shared digital board.

7. Closing and Next Steps

- Thank participants and recap key points discussed.
- Highlight how their feedback will shape future developments in marking practices.

Appendix C

Expert Semi Structured Interview

Phase 1

First Element - Your CJ Practice

- Overview of your current CJ practice.
- Key methodologies and approaches used.
- Challenges faced in implementing CJ.
- Effectiveness of CJ in assessment.
- Potential improvements and refinements.

Second Element - Transparency and Reliability of CJ

All elements are considered from:

- Student perspective
- Educator perspective

Key aspects of transparency and reliability:

- Transparency of the mark derived – how rank informs the mark and how rank is decided.
- How student work informs the rank that is derived – the elements of it that influence it.

- Reliability of the markers doing the work and following the rules.
- Transparency of the comparison selection process (random vs. Bayesian).

Third Element - BCJ and MBCJ for Transparency

- Explanation of BCJ and its role in assessment.
- How BCJ enhances transparency in the judgement process.
- Introduction to Multi-Criteria BCJ and its advantages.
- Comparison of BCJ and Multi-Criteria BCJ in terms of reliability and fairness.
- Challenges and potential improvements in implementing BCJ-based systems.

Phase 2

Fourth Element - Presentation and Reflection on Results

Key considerations:

- Are the results convincing?
- Are the results comprehensible?
- Do they address the questions that the expert has about MBCJ?

Fifth Element - Future Design

What further data could be gathered to support the MBCJ concept?

Achieving transparency outside of research settings if τ scores cannot be used:

- Seeded judgments where the lead knows the answer.
- Identifying outlier marking.

Appendix D

Open Source BCJ Python Library

As a result of the research carried out in this thesis, a Python open source project has been created allowing users to carry out pair-wise comparisons using either BTM, BCJ or MBCJ. The PYPI installation can be found here: https://pypi.org/project/comparative-judgement/, while the GitHub repository can be found here: https://github.com/codingWithAndy/comparative_judgement.

Appendix E

BCJ Web App

The URL for the BCJ Web Application https://github.com/codingWithAndy/BayesCJ-Web-App

Appendix F

MBCJ Web App

The URL for the multi-criterion BCJ Web Application https://github.com/codingWithAndy/BayesCJ-multi-dimensional-Web-App