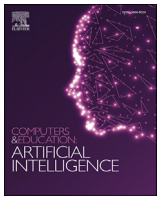
Computers and Education: Artificial Intelligence 6 (2024) 100245

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence



A Bayesian active learning approach to comparative judgement within education assessment

Andy Gray [a,*], Alma Rahat [a], Tom Crick [a], Stephen Lindsay [b]

[a] Swansea University, Swansea, United Kingdom
[b] University of Glasgow, Glasgow, United Kingdom

ARTICLE INFO

Keywords:
Comparative judgement
Bayesian learning
Active learning
Machine learning
Assessment
Bradley-Terry model (BTM)

ABSTRACT

Assessment is a crucial part of education. Traditional marking is a source of inconsistencies and unconscious bias, placing a high cognitive load on the assessors. One approach to address these issues is comparative judgement (CJ). In CJ, the assessor is presented with a pair of items of work, and asked to select the better one. Following a series of comparisons, a rank for any item may be derived using a ranking model, for example, the Bradley-Terry model, based on the pairwise comparisons. While CJ is considered to be a reliable method for conducting marking, there are concerns surrounding its transparency, and the ideal number of pairwise comparisons to generate a reliable estimation of the rank order is not known. Additionally, there have been attempts to generate a method of selecting pairs that should be compared next in an informative manner, but some existing methods are known to have created their own bias within results inflating the reliability metric used within the process. As a consequence, a random selection approach is usually deployed.

In this paper, we propose a novel Bayesian approach to CJ (which we call BCJ) for determining the ranks of a range of items under scrutiny alongside a new way to select the pairs to present to the marker(s) using active learning, addressing the key shortcomings of traditional CJ. Furthermore, we demonstrate how the entire approach may provide transparency by providing the user insights into how it is making its decisions and, at the same time, being more efficient. Results from our synthetic experiments confirm that the proposed BCJ combined with entropy-driven active learning pair-selection method is superior (i.e. always equal to or significantly better) than other alternatives, for example, the traditional CJ method with differing selection methods such as uniformly random, or the popular no repeating pairs where pairs are selected in a round-robin fashion. We also find that the more comparisons that are conducted, the more accurate BCJ becomes, which solves the issue the current method has of the model deteriorating if too many comparisons are performed. As our approach can generate the complete predicted rank distribution for an item, we also show how this can be utilised in probabilistically devising a predicted grade, guided by the choice of the assessor. Finally, we demonstrate our approach on a real dataset on assessing GCSE (UK school-level) essays, highlighting the advantages of BCJ over CJ.

1. Introduction

The core mathematical technique used for generating ranks from paired comparisons in comparative judgement for assessment was originally proposed in 1927 (Thurstone, 1927). In this paper, for the first time, we propose a Bayesian approach, appropriately considering the epistemic uncertainty arising from limited number of comparisons, and propagating it through to the 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 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.

Subjectivity, bias and inequity influence the overall judgement on a pupil's performance (Finn & Cinpoes, 2022); leading to asking fundamental questions such as: is assessment fair? (Nisbet & Shaw, 2020).

* Corresponding author.
E-mail addresses: 445348@swansea.ac.uk (A. Gray), a.a.m.rahat@swansea.ac.uk (A. Rahat), thomas.crick@swansea.ac.uk (T. Crick), stephen.lindsay@glasgow.ac.uk (S. Lindsay).

https://doi.org/10.1016/j.caeai.2024.100245
Received 12 September 2023; Received in revised form 21 May 2024; Accepted 22 May 2024
Available online 27 May 2024

For example, inconsistency in teachers predicting student grades is widespread in UK schools and colleges. In 2019, only 21% of students obtained the grades predicted by their teachers (Jeffreys, 2022) while in 2011, 42-44% of teacher estimated grades over-predicted by at least one grade, and 7-11% under-predicted (Everett & Papageorgiou, 2011). The COVID-19 pandemic forced reliance on predicated grades across educational settings and contexts. The impact was immediate and profound (Watermeyer, Crick, et al., 2021; Crick et al., 2020; Marchant et al., 2021; Crick et al., 2021; Siegel et al., 2021; Lowthian et al., 2023) and its long-term consequences are still not fully manifested (Watermeyer, Shankar, et al., 2021; Shankar et al., 2021; McGaughey et al., 2022; Hardman et al., 2022); we will likely continue to experience a "new normal" for education for some time (Crick, 2021; Ward et al., 2021; Watermeyer et al., 2022a; Irons & Crick, 2022; Crick, Knight, et al., 2022; Thomas et al., 2023), and especially for educational assessment (Watermeyer et al., 2022b; Siegel et al., 2021; Crick, Prickett, et al., 2022; Ward et al., 2023; Knight et al., 2023). During the pandemic, student grades were given 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 As and A^* results, a fall from 19.1% getting A^*s in 2021 compared to 13.5% in 2022 – ultimately bringing grades back in line with pre-pandemic results (Weale, 2022).

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 (Luckin et al., 2016; Namoun & Alshanqiti, 2020; Rastrollo-Guerrero et al., 2020; Dwivedi et al., 2021; Shafiq et al., 2022); in particular, for predicting student performance and retention we have seen broad application of data mining and learning analytics (Elbadrawy et al., 2016; Yağcı, 2022), as well as machine learning, collaborative filtering, recommender systems, and artificial neural networks (Iqbal et al., 2017; Vijayalakshmi & Venkatachalapathy, 2019; Yousafzai et al., 2020). However, there are increasingly complex and interconnected social, ethical, legal and digital/data rights issues with these varied approaches (Slade & Prinsloo, 2013; Williamson, Bayne, et al., 2020; Akgun & Greenhow, 2019), especially with pre- and post-pandemic critical analysis (Williamson, Eynon, et al., 2020). 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.

Prospect theory shows that humans are better at identifying relative, rather than absolute quality (Kahneman & Tversky, 2013). In educational assessment, this has been recognised (Benton & Gallagher, 2018), and comparative judgement (CJ) is an alternative to traditional marking (Pollitt & Murray, 1996). In CJ, an assessor is presented with pairs of work, and they only decide which is of higher quality instead of assigning an absolute mark. The process is repeated a predefined number of times, potentially re-evaluating pairs. A ranked order of items is then derived from these pairwise comparisons using a model of CJ, for example, the Bradley-Terry model (Hunter, 2004), which was inspired by Thurstone's mathematical definition of ranking from comparisons (Thurstone, 1927). In this way, we are able to extract an accurate ranked order from only a series of relative comparisons. In addition, an important benefit of CJ is that the cognitive load placed on the teachers while marking is also reduced (Coenen et al., 2018).

However, one of the key drawbacks of CJ is that, irrespective of specific approaches, it can take numerous iterations (that is, the number of pairs to be assessed) and significant time to complete the marking, in addition to the time required to collate grades, award students' scores, and then provide feedback. Alternative CJ methods, for example adaptive comparative judgment (ACJ), are designed to reduce interactions without loss of accuracy, but have been found to include other bias through their "adaptive nature" (Bramley, 2015). Therefore, the pure form is still the desired version. This means that, although CJ has its advantages, there is still a significant research problem in finding a method that can decrease the number of interactions and the overall time required for marking.

Furthermore, Ofqual, the official governmental body that regulates qualifications, exams, and tests in England, has also pointed out that CJ's paired comparison rank order starts to deteriorate, and the entire model's accuracy begins to deteriorate unless it is precisely determined in advance what the minimum number of judgments required is with a level of confidence that is currently unknown (Holmes et al., 2020). Furthermore, Ofqual also holds the belief that CJ faces issues regarding its lack of transparency in the manner it formulates and presents its conclusions (Holmes et al., 2020).

We thus propose a novel Bayesian approach towards 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 paper 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 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 Bramley and Vitello (2019), and highlight the advantages of the proposed method in comparison to standard CJ.

The rest of the paper is structured as follows: in Section 2, we present related work and some background; Section 3 outlines how the main algorithms work to rank students' work. We will explain the three methods used to select the next pairs to be compared in Section 4; we present our results and discussions in Section 5, with general conclusions and future work in Section 7.

2. Related work in education

CJ is a technique used to derive ranks from pair-wise comparisons. The concept of CJ is used in academic settings to allow teachers to compare two pieces of work and select which is better against selected criteria. After each comparison, another pair is selected. This is repeated until enough pairs have been compared to generate a ranking of the work marked. We detail a typical CJ process in Algorithm 1.

An important benefit to CJ within an academic setting is reducing the teacher's cognitive load (Chen et al., 2023), as comparing two pieces of work is faster than marking each individual piece, while also insisting the teacher is being non-biased towards a student and consistent Sadler (1989). This is difficult to achieve (Bramley, 2007), and CJ helps, to an extent, address this challenge; for a further discussion of this, we refer to the following literature (Benton & Gallagher, 2018; Bartholomew et al., 2019; Christodoulou, 2017).

Algorithm 1 Standard comparative judgement procedure.

Inputs.

N: Number of items.
K: Multiplier to calculate the budget for the number of pairs to be assessed.
I: Set of items.

Steps.

1: $B \leftarrow N \times K$ ▷ Compute the budget.
2: $G \leftarrow \langle\rangle$ ▷ Initialise list of selected pairs.
3: $W \leftarrow \langle\rangle$ ▷ Initialise list of winners.
4: $\mathbf{r} \leftarrow \left(\frac{N}{2}, \ldots, \frac{N}{2}\right)^\top \mid |\mathbf{r}| = N$ ▷ Initialise rank vector with mean rank for all items.
5: **for** $b = 1 \rightarrow B$ **do**
6: $(i, j) \leftarrow \text{SelectPair}(I)$ ▷ Pick a pair of items.
7: $G \leftarrow G \oplus \langle(i, j)\rangle$ ▷ Append the latest pair.
8: $w \leftarrow \text{DetermineWinner}(i, j)$ ▷ Pick a pair of items.
9: $W \leftarrow W \oplus \langle w\rangle$ ▷ Append the latest winner.
10: $\mathbf{r} \leftarrow \text{GenerateRank}(G, W)$ ▷ Update rank vector.
11: **end for**
12: **return r**

CJ is based on a technique originally proposed by Thurstone in 1927, known as 'the law of comparative judgement' (Thurstone, 1927). Thurstone discovered that humans are better at comparing things to each other rather than making judgements in isolation, for example, judging if a piece of fruit is bigger than another without having the other fruits to compare against at the point of judgement. Therefore, he proposed making many pair-wise comparisons until a rank order was created (Thurstone, 1927; Benton & Gallagher, 2018; Bartholomew et al., 2019). Pollitt et al. introduced and popularised it within an education setting (Pollitt & Murray, 1996; Pollitt, 2004).

Typically, the efficacy of a CJ method is measured using the Scale Separation Reliability (SSR) (Bramley, 2015; Pinot de Moira et al., 2022; Pollitt, 2012). SSR is defined as the ratio between the variance of the true score and the variance of estimated scores from observations; interested readers should refer to the work of Verhabert et al. (Verhavert et al., 2018) for a detailed discourse on SSR. The relative uncertainty estimation through SSR is highly dependent on the underlying CJ model (e.g. BTM) and its own estimated uncertainty, which is typically not presented to the users of the system in an intuitive way. SSR might not even be calculable, as it requires knowledge of variance of true scores, which is not available in most practical cases.

An important consideration in CJ is the stopping criterion. To the best of our knowledge, there seems to be no natural and meaningful performance metric that would allow for a clear indication on when to stop. Because of this, CJ is usually conducted on a fixed budget, giving the number of pairs that must be compared before finalising the rank order, for example, at least 10 judgements per script (Wheadon et al., 2020).

A growing body of evidence supports the use of CJ as a reliable alternative for assessing open-ended and subjective tasks. Teachers' judgements, more generally termed *raters* or *judges*, are fed into a BTM (see Section 3.1 for more details on the BTM) to produce scores that represent the underlying quality of the scripts (Bradley & Terry, 1952; Luce, 1959). These scores have the appealing property of being equivalent across comparisons (Andrich, 1978).

A key justification for using CJ within the educational assessment process is that the rank orders it produces tend to have high levels of reliability. For example, in 16 CJ exercises conducted between 1998 and 2015, the correlation coefficient scores were between 0.73 to 0.99 when compared with rubric-based grades (Steedle & Ferrara, 2016). With a correlation coefficient of 1.0 representing perfect agreement, a score of 0.70 or greater is typically considered high enough to declare strong agreement (Hinkle et al., 2002).

Alternative methods for conducting CJ, such as ACJ, have been introduced. These methods differ in terms of how the pairs to be evaluated next are selected or assigned to assessors. ACJ is a version that aims to be adaptive based on the current state of the marking between the judges. The adaption is based on an algorithm that pairs items ranked similarly as the judge progresses in the CJ process, a method aimed at expediting the process of achieving an acceptable level of reliability (Bartholomew et al., 2019). Pollitt first proposed ACJ in 2011, a system created in partnership with TAG assessments (Pollitt, 2012). Later, the system was further developed by RM Compare (Jones & Davies, 2022).

A significant flaw in the ACJ approach was that its adaptive nature generated its own bias by having more similar pieces of student work compared to themselves more frequently, and thus the correlation between true reliability and SSR (due to ACJ) has been shown to be low in some experiments (Bramley, 2015). Further, the process usually takes longer than traditional marking (Benton & Gallagher, 2018; Bramley, 2015). Therefore, it is suggested that having random pairings is just as effective as the ACJ approach. As a result, the CJ community has reverted, to some extent, to random pairings and removed the adaptive nature of the CJ process (Wheadon et al., 2020; Jones & Davies, 2022).

Furthermore, claims have been made that CJ advocates have no compelling case to support two of their central claims: that humans are better at comparative than absolute judgments and that CJ is necessarily valid because it aggregates judgments made by experts in a naturalistic way (Kelly et al., 2022). However, there are experiments that provide clear evidence of human efficiency in CJ in general (Kahneman & Tversky, 2013), and the practical consistency of CJ for marking (Steedle & Ferrara, 2016). However, there is a lack of clarity in how the decisions are being made, and we note this as one of the key criticisms of CJ, along with the lack of estimations of uncertainty in estimations, despite the practical strengths. Our investigation in BCJ was primarily driven by these criticisms with the aim of improving the state-of-the-art of CJ.

It should be noted that much attention has been placed within education on ensuring learning standards (Bloxham & Price, 2015). Learning Standards is the fixed level of achievement expected of a student to be awarded a recognised grade (O'Connell et al., 2016). A key aspect to successfully implementing a shared rubric is the process of norming, also known as calibrating or moderating rubrics (Schoepp et al., 2019). Norming is a collaborative process built around knowledge of the rubric and meaningful discussion leading to evidence-driven consensus (Schoepp et al., 2019). Depending on the time-frames available, it is preferable also to have calibration sessions in which experts discuss the scoring rules with each other (Wammes et al., 2022).

CJ system is built around getting multiple markers to make multiple judgements on presented pieces of work (Benton & Gallagher, 2018; Leech & Chambers, 2022). Studies on multiple marking for CJ has favoured positive reviews (Benton & Gallagher, 2018), therefore claiming that CJ can be used as a form of moderation technique between markers. Which, as a result, enables the calibration of moderated work within its process (Jones & Davies, 2022; Elander & Hardman, 2002). Any improvement that we make to CJ would inherit this feature by extension.

In the following section, we will first focus on describing the generation of ranks from paired comparisons (in line 10 of Algorithm 1), as the pair selection method (in line 6 of Algorithm 1) we propose depends on the concepts required for rank generation.

3. Generating ranks from a list of paired comparisons

Currently, the most popular method of ranking paired comparisons is BTM. Therefore, in this section, we will first explain how the BTM system works and then provide a description of our proposed Bayesian approach.

3.1. Classical approach: Bradley-Terry model

Bradley and Terry proposed BTM in their seminal paper on the topic (Benton & Gallagher, 2018; Bisson et al., 2016; Marshall et al., 2020; Pollitt, 2012; Gray et al., 2022). Traditionally, this has been adopted as the driving algorithm for CJ. The technique is an iterative minorisation-maximisation (MM) method (Hunter, 2004) to estimate the maximum likelihood of the expected preference score γ_i for the ith student's item of work, given the observed data. With the expected preferences, we can then use this to arrange the items of work and then generate a rank where a higher value represents a better quality of work. We present a mathematical description of the model below, broadly following Hunter's work (Hunter, 2004).

Consider the set of N items, $I = \{1, \ldots, N\}$ with each element i representing the identifier of the relevant item. The expected performance vector is $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N)^\top$, where γ_i is a positive parameter representing the overall score for the ith item. For example, in a typical marking context, we can assume that an individual's mark varies between 0 and 100, i.e. $\gamma_i \in [0, 100]$; however, this assumption is not essential for the scheme to work, and thus can be safely ignored. Now, the probability that the ith item is of higher quality compared to the jth item is given by:

$$P(i \succ j) = \frac{\gamma_i}{\gamma_i + \gamma_j}. \tag{1}$$

Using the key assumption that the outcomes of different pairings are independent, the log-likelihood for the performance vector $\boldsymbol{\gamma}$ is given by:

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

where $\omega_{[i,j]}$ is the number of times item i was preferred over item j. It should be noted that typically BTM ignores any notion of ties, and raters are forced to make a decision on the winner.

The minorisation-maximisation (MM) algorithm proposed by Hunter (Hunter, 2004) iteratively updates each γ_i such that the log-likelihood in (2) is maximised. The iterative update formula for kth iteration is (Gescheider, 2013):

$$\gamma_i^{k+1} = \Omega_i \sum_{j|j\neq i} \frac{\omega_{[i,j]} + \omega_{[j,i]}}{(\gamma_i^k + \gamma_j^k)} \tag{3}$$

Where, $\Omega_i = \sum_j \omega_{[i,j]}$ is the number of times the ith item was preferred. At each iteration, we are further required to normalise the γ_is to ensure that the sum of the elements of the performance vector equals 1, i.e.

$$\gamma_i^{k+1} \leftarrow \frac{\gamma_i^{k+1}}{\sum_j \gamma_j^{k+1}}. \tag{4}$$

Under certain assumptions, the iterative process will converge to the optimal $\boldsymbol{\gamma}$ (Hunter, 2004). In this work, at the final stage, for ease of presentation and assuming $\gamma_i \in [0, 1]$, we multiply γ_i by 100. We can then extract the rank of the ith item as follows (using 1-based counting):

$$r_{i\in I} = (N + 1) - \text{argsort}(\boldsymbol{\gamma}). \tag{5}$$

The process in Equation (5) can be repeated to generate the complete rank vector in line 10 of Algorithm 1.

3.2. Proposed Bayesian approach

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 (2) 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. (2021) for a concise and recent overview, and to McElreath (2020), or Lambert (2018), 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, and produces similar results to standard CJ in terms of identifying the ranking (Pritikin, 2020; Wainer, 2022; Tsukida & Gupta, 2011; De Maeyer, 2021). 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 (Wainer, 2022)), 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 4.3.

3.2.1. 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 (Sivia & Skilling, 2006):

$$L(p|\mathbf{x}) \propto p^w (1-p)^{n-w}. \tag{6}$$

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 (Fink, 1997):

$$\alpha \leftarrow \alpha_{init} + w, \tag{7}$$

$$\beta \leftarrow \beta_{init} + (n - w). \tag{8}$$

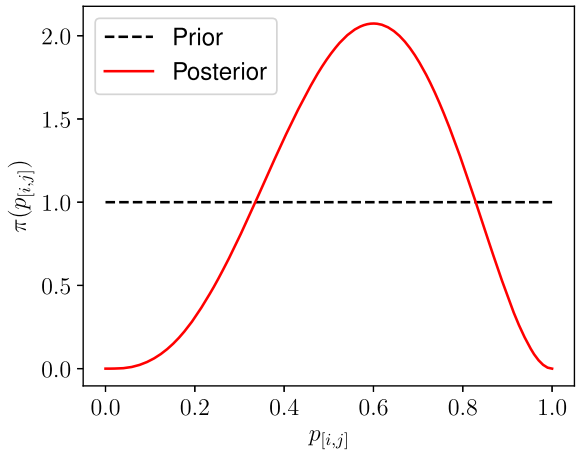


Fig. 1. A toy example of Bayesian updating of PDF over preference between ith and jth items. Initially, with uniform prior (shown with a black dashed line), none is preferred. Then, with three wins ($\alpha = 1 + 3 = 4$) and two losses ($\beta = 1 + 5 - 3 = 3$) for i after five comparisons, the PDF (depicted 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 would reduce.

With priors of $\alpha_{init} = 1$ and $\beta_{init} = 1$, we get a uniform prior, as in 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 Fig. 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.

With this framework, we define the probability that i is preferred over j, i.e. a different interpretation of probability of winning in (1), as:

$$P(i \succ j) = P\left(\pi(p_{[i,j]}) > 0.5\right) = 1 - \mathcal{F}(0.5), \tag{9}$$

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:

$$P(j \succ i) = 1 - P(i \succ j). \tag{10}$$

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

3.2.2. 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 (10), 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 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):

$$P(r_i = 1) = \prod_{j \in I \setminus \{i\}} P(i \succ j). \tag{11}$$

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

$$P(r_i = N) = \prod_{j \in I \setminus \{i\}} P(j \succ i). \tag{12}$$

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 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:

$$P(r_i = a | D_k, E_k) = \prod_{s \in D_k} P(s \succ i) \prod_{t \in E_k} P(i \succ t). \tag{13}$$

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

$$P(r_i = a) = \sum_{k=1}^{z_a} P(r_i = a | D_k, E_k), \tag{14}$$

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 (Feller, 1968):

$$\mathbb{E}[r_i] = \sum_a a P(r_i = a). \tag{15}$$

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 419m 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 this expense of computing the expected rank of an item in (15) is to use a form of numerical integration. In fact, a simple Monte Carlo (MC) integration (Mackay, 1998) 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 $r'_i = (N + 1) - w'$; cf. with (5). For R samples, we can then estimate the expected rank of i as follows:

$$\mathbb{E}[r_i] = \frac{1}{R} \sum_{k=1}^{R} r'_i[k], \tag{16}$$

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 (Koehler et al., 2009). 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. So, in this case, we would need $10000N$ samples to estimate ranks for all items, which can be done efficiently in a standard desktop computer, even for large N.

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

$$r_{i \in I} = (N + 1) - \text{argsort}(\mathbb{E}[\mathbf{r}]). \tag{17}$$

We present an illustrative synthetic example in the following section.

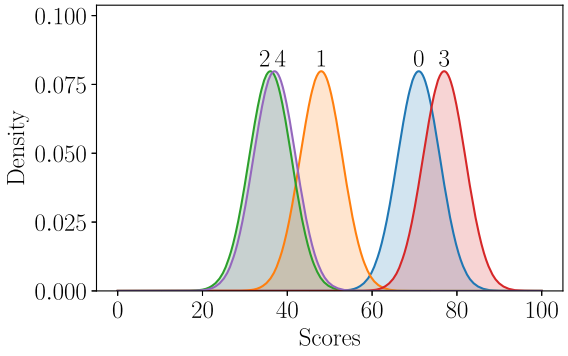


Fig. 2. An illustration of five items with Normally distributed scores. Here, the mean vector for the items is $\boldsymbol{\mu} = (71, 48, 36, 77, 37)^\top$ and $\sigma = 5$ to represent uncertainty around the mean scores. The σ also represents the range of marks multiple judges could give the piece of work from traditional 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 whichever yields the higher score wins.

3.2.3. An illustration

We consider a set of *five* items with respective scores and the associated uncertainties, as shown in Fig. 2. We assume that the scores are Normally distributed (as per 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 paper 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 (14) as we can calculate the probability that i dominates j as follows (Hughes, 2001):

$$P(i \succ j) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{m}{\sqrt{2}}\right)\right], \tag{18}$$

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 $\operatorname{erf}(\cdot)$ represents the Gauss error function (Andrews, 1998).

In Fig. 3, we show the target distribution over ranks for the items in Fig. 2, calculated using (14) and (18). 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 (Thiagarajan & Ghosh, 2023) 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 get 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 instead to rank the items. To compare our approach with BTM based CJ, we will therefore use the BCJ expected ranks to identify the ranks of items.

In Fig. 4, 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 paper, 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.

4. Selecting a pair of items to compare

One of the key questions when implementing a CJ approach for marking is how to select the next pair to evaluate (step 6 in Algorithm 1) to identify comparative preference. There are many ways to generate these pairs, see, for example, Jones and Davies (2022), but these are typically *ad hoc* in nature. Furthermore, Ofqual has stated that if the number of pairs goes too far over the optimal number, then the final ranking becomes less effective, but knowing this optimal number of comparisons is unknown (Holmes et al., 2020). Although CJ is typically fast and offers a good means of ranking items of work, it does not give insight into how the model generated its results.

Our goal in this paper is to provide further insight into the process for the assessors, particularly the uncertainties illustrated in the previous section. More importantly, we want to drive the selection of the pairs to be evaluated using the knowledge that we have already gathered, and thus facilitate decision-making in an informed manner to reduce the need for many evaluations.

It should be noted that the traditional stopping criterion is usually expressed as a budget on the number of pairs evaluated: here, we assume that the budget is $N \times K$ where K is the multiplier that is often set to 10 (Jones & Davies, 2022).

In this section, we describe three ways to identify the next pair to be compared: randomly, using NRP and our novel entropy approach.

4.1. Random

The random approach picks every pair presented to the user at random until the budget is reached. This can cause the same pair to be presented to the user, but that would be unlikely, especially as N increases in size. This is effectively a random search method that is known to be effective for high-dimensional problems (Bergstra & Bengio, 2012). We use this widely used method (Jones & Davies, 2022; Benton & Gallagher, 2018) as a baseline for comparison.

4.2. No repeating pairs

This is another approach used within current approaches: it is a round-robin approach, where no repeating pairs occur until we have selected all possible pairs (Jones & Davies, 2022; Holmes et al., 2020). This ensures that all N items are seen the same number of times, but what item is compared against what item is decided uniformly as random. This prevents the same pairs from being presented to a user until every other pair has been rated. However, as we have no indication of uncertainty, certain pairs may be selected despite the difference between them being clear.

4.3. Active learning with entropy

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 (Settles, 2010; Knijnenburg & Willemsen, 2015; Das et al., 2016). 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 (MacKay,

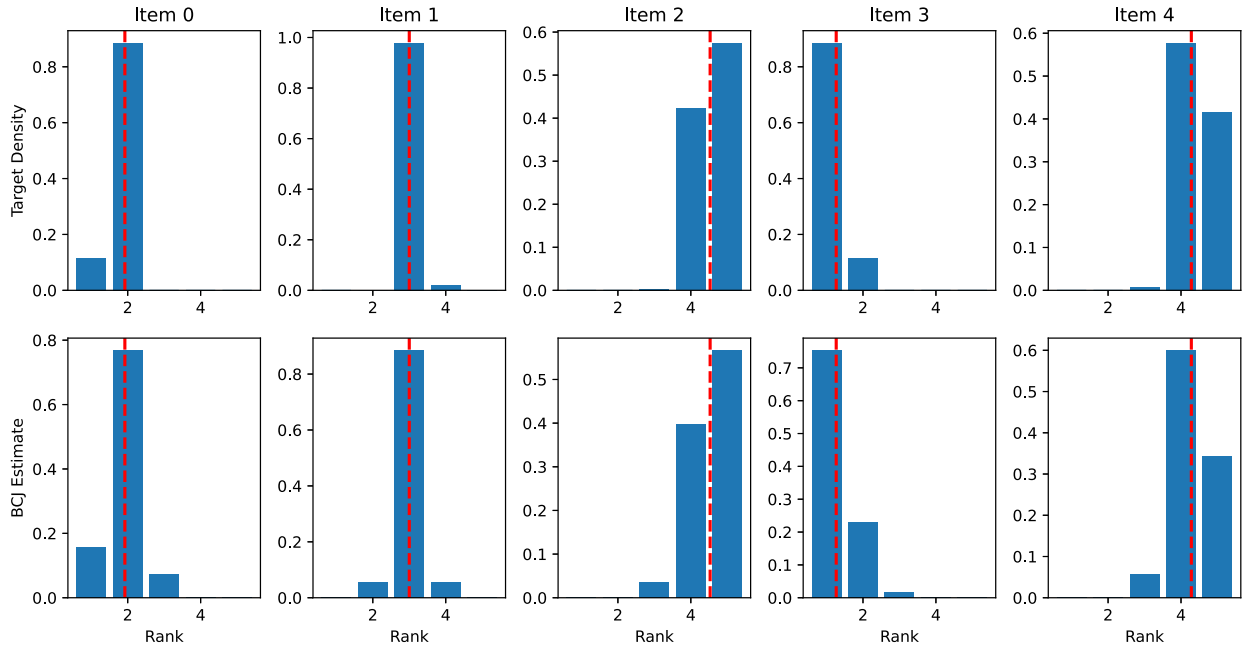


Fig. 3. Probability distributions of ranks of items presented in Fig. 2. The top row shows the densities calculated directly from the Normal distributions over the scores using (14). The bottom row shows the estimated rank distributions using our proposed BCJ method after $N \times K = 5 \times 10 = 50$ pairwise comparisons (driven by our entropy based active learning method presented in Section 4.3). 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 $\mathbb{E}[\mathbf{r}]$.

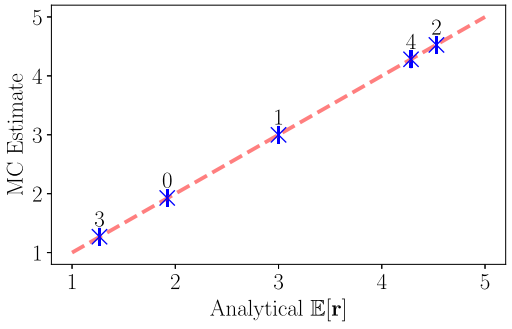


Fig. 4. Comparison between the analytical estimates in (14) and Monte Carlo estimates (with 10k samples) in (16) of the expected rank vector of the items $\mathbb{E}[\mathbf{r}]$ for our proposed BCJ method after $N \times K = 5 \times 10 = 50$ comparisons as in Fig. 3. Crosses show the mean MC estimate, 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.

1992). In this way, we collect data efficiently and learn a good model with fewer data points.

There are many variants of AL. In this paper, we focus on so-called "pool-based learning" (Zhan et al., 2021) 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 (Lewis, 1995).

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 being represented by higher entropy, and for the posterior Beta density of BCJ, it can be computed as (Lazo & Rathie, 1978):

$$H\left[\pi(p_{[i,j]})\right] = \ln B(\alpha,\beta) - (\alpha-1)\psi(\alpha) - (\beta-1)\psi(\beta) + (\alpha+\beta-2)\psi(\alpha+\beta), \tag{19}$$

where, $B(\alpha,\beta)$ is the Beta function and the $\psi(\cdot)$ is the Digamma function. Given the parameters α and β, this is straightforward to calculate using existing statistics packages, for example, `scipy.stats` in Python (Virtanen et al., 2020).

In this paper, 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.

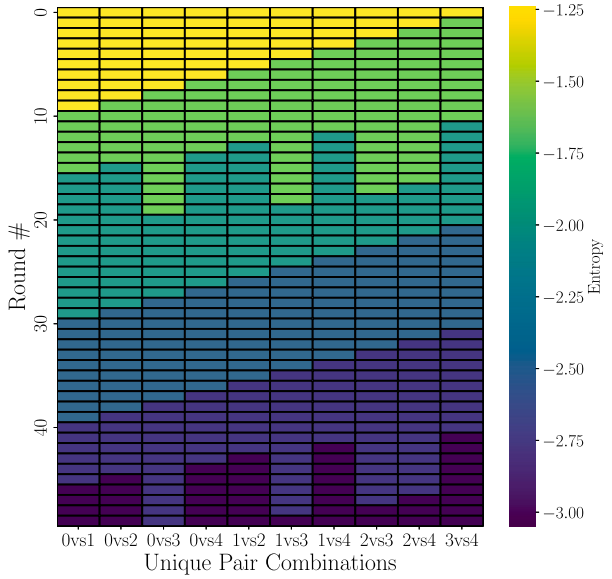
In Fig. 5, we demonstrate the entropy score after each round of comparisons, and the associated selection process. The process involves the algorithm calculating the entropy value for each pair combination in $\mathcal{P}$ to see which pair has the highest value, and then selecting that pair to be presented. 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 Fig. 2.

5. Experiments and discussions

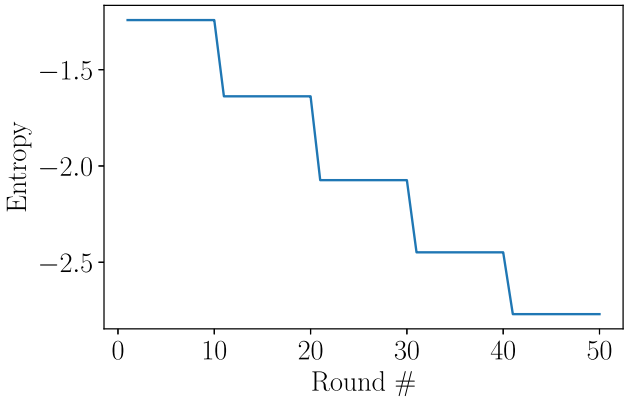
In this section, we will present our findings, analyse them, and discuss what we believe they represent and mean.

In our reading of the literature, we found that the suggested budget for the number of comparisons were $N \times K = 10N$ (Jones & Davies, 2022). 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 3.2.3); these



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



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

Fig. 5. Illustration of uncertainty sampling using entropy (*top*) for the five items in Fig. 2 after $N \times K = 50$ comparisons, and the respective gradual reduction in maximum entropy (*bottom*). As a pair is selected, its uncertainty reduces immediately after data is gathered about preference. The downward trajectory in maximum entropy shows that the model is becoming more accurate over iterations.

were paired, and therefore, we performed Wilcoxon Rank-Sum tests on the final results with Bon-Ferroni correction for multiple comparisons (Miller, 1981) at a significance level of $\alpha = 0.05$.

Measuring the performance of the methods is not straightforward. We consider that the targets of the scores of items have uncertainty and are Normally distributed. Traditional CJ only generates a single rank for items without any uncertainty. To compare the results, we use the target distributions to derive the expected rank of each item, and then sort the items by expected rank, giving a target rank; see Equation (17). This allows us to measure performance via normalised Kendall's τ rank distance, which measures the difference between two ranking lists. The metric is calculated by counting the discrepancies between the two lists. The greater the distance, the more disparate the lists (Kendall, 1938; Fagin et al., 2003). The normalised distance ranges from 0 (indicating perfect agreement between the two lists) to 1 (indicating complete disagreement between the lists). For example, a distance of 0.03 means that only 3% of the pairs differ in ordering. In this paper, when a method progressed, we noted the τ distance after each paired comparison, and this showed how well the relevant method converged to the target rank.

It should be noted that BCJ can estimate the whole distribution. So, we can compute JSD, as discussed in Section 3.2.3, to identify the agreement between target and BCJ estimated densities.

In the following sections, we first discuss the performance of different methods in terms of τ distance. Then we discuss how well BCJ does in estimating the complete target distributions in terms of JSD. Finally, we propose yet another method for assigning grade letters to individual items based on the complete probability distribution over the rank of an item.

5.1. Analysing the winning method

In Fig. 6, 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 (Holmes et al., 2020), 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 Fig. 7. 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 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.

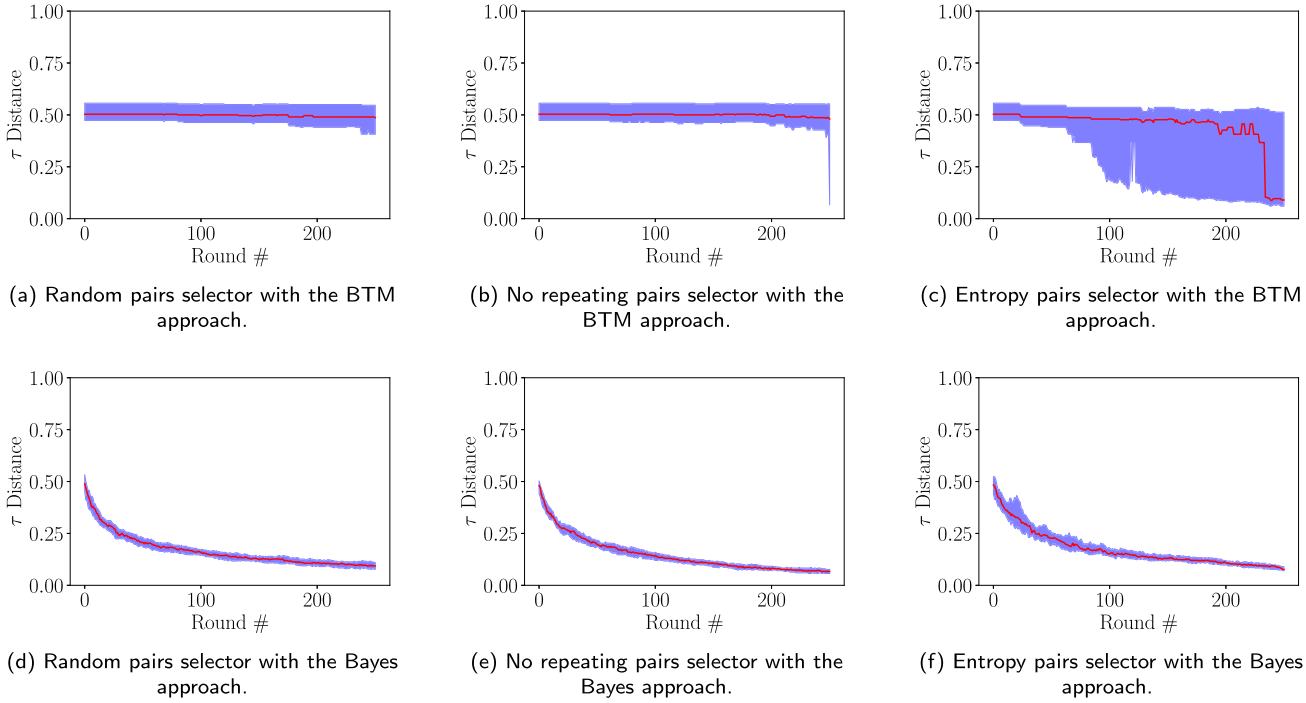
The count of times that a method i has been beaten by other methods can be calculated with the following expression: $V(i) = \sum_{i \neq j \wedge j \in S} \left[\text{p-value}(i > j) \leq \alpha_{adj}\right]$, where p-value$(i > j)$ is the binary result of comparing i and j with 1 representing that i has a statistically higher value than j (as in i is worse than j in a one-sided manner), and the adjusted significance level is defined as $\alpha_{adj} \leftarrow \frac{\alpha}{m}$, with the original significance level $\alpha = 0.05$ and the number of comparisons $m = 5$ for every combination where these tests were performed, using Bon-Ferroni correction for multiple comparisons (Miller, 1981).

These results are shown in Fig. 8. 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 Fig. 8f 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 Figs. 6 and 7 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.

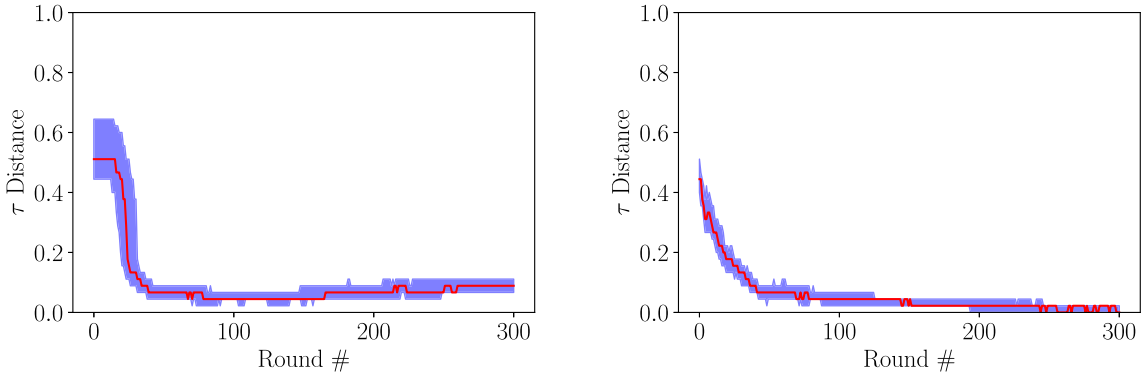
Therefore, overall we can suggest that the Bayes version as a ranking method has done better, but the combination of Bayes and Entropy has done the best overall. Especially when comparing the current state-of-the-art approach (Fig. 8b) and our two novel approaches (Fig. 8f).

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 Fig. 1 for investigating



(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.

Fig. 6. A comparison of the random (6a, 6d), no repeating pairs (6b, 6e) and entropy (6c, 6f) 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.



(a) No Repeating pairs selector with the BTM approach.
(b) Entropy pairs selector with the Bayes approach.

Fig. 7. Convergence plots of the main current method for conducting CJ, a combination of the NR pairing method and BTM (Fig. 7a), and our novel entropy pairing method with BCJ (Fig. 7b). 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.

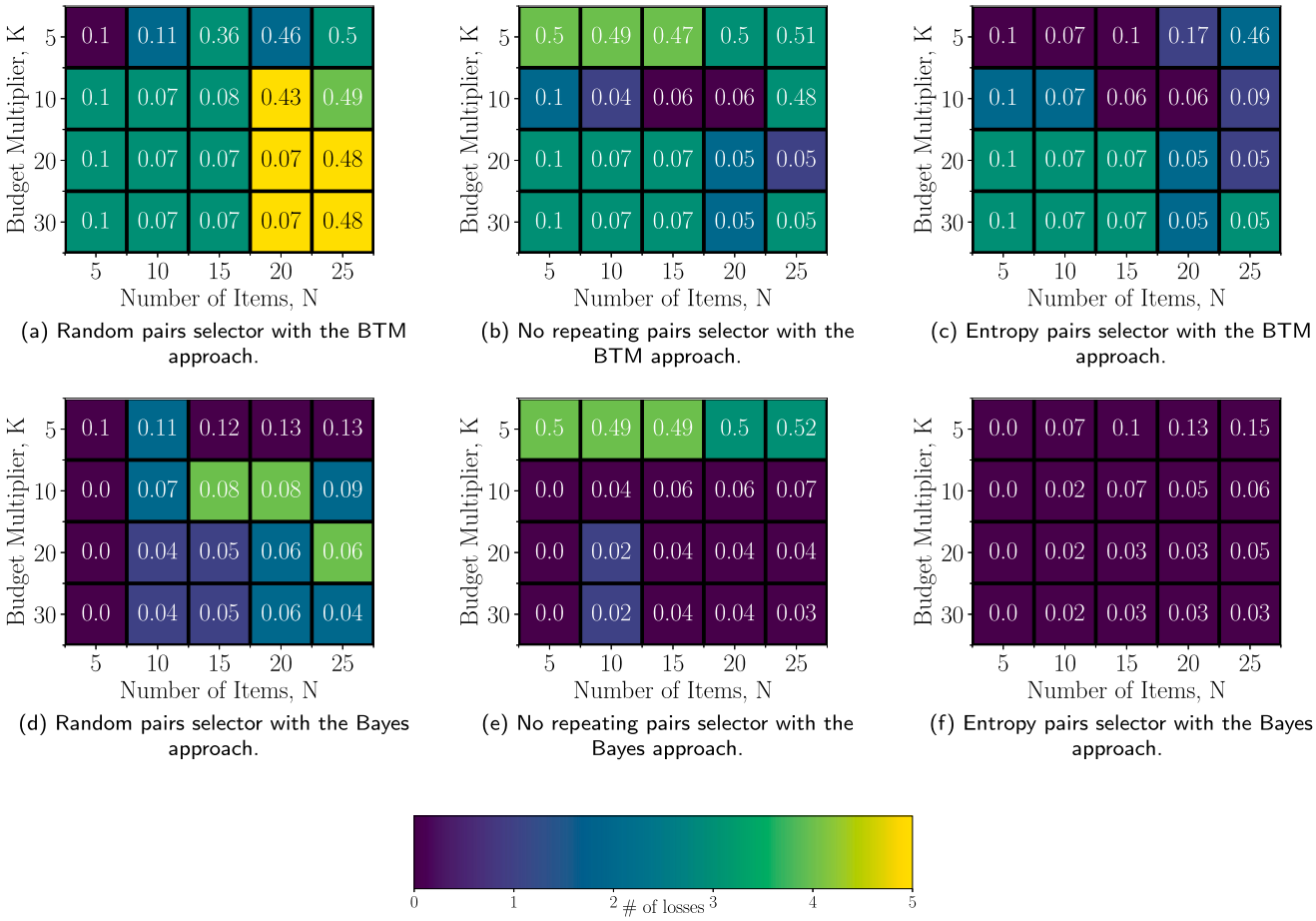
the current state of the preference PDF between any pair of items, and deriving the resulting rank distribution in Fig. 3 (bottom row). One can also track the entropy reductions using Fig. 5.

5.2. 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*.

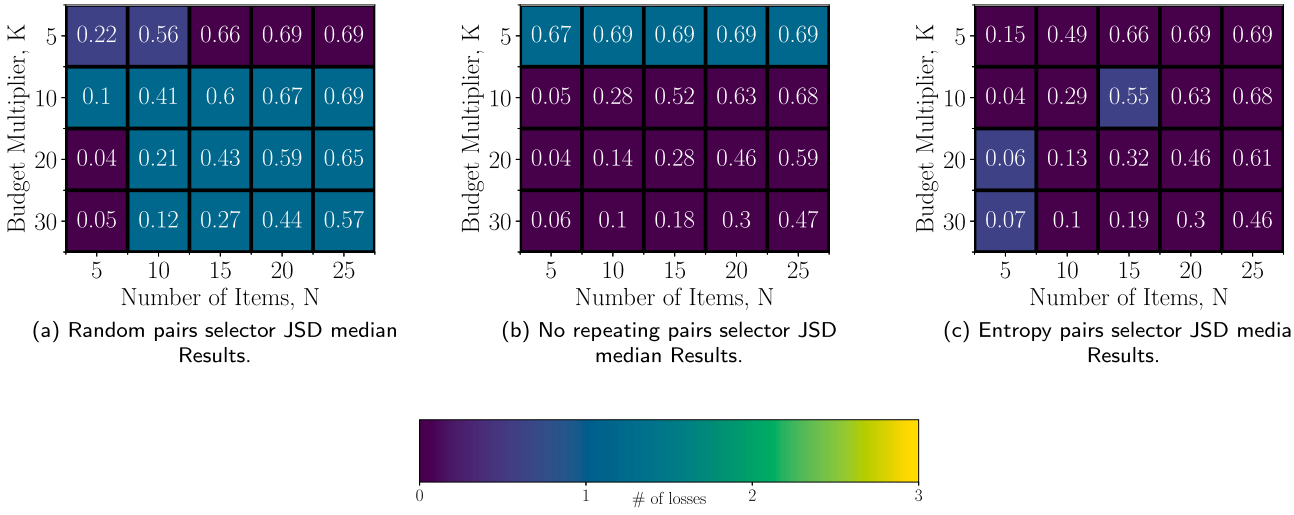
Here, we use the JSD measure to identify the agreement between our BCJ estimate and the actual target distributions. For N items, we deduce N distributions over ranks and compare with its target counterpart. This comparison gives us N JSD values. We take the worst JSD as reflective of the performance of the current rank distribution and track this throughout the BCJ process as a measure of progress.

The results in Fig. 9 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



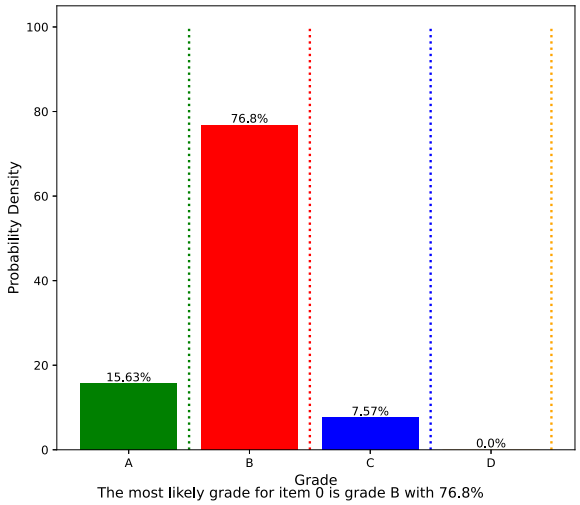
(a) Random pairs selector with the BTM approach.
(b) No repeating pairs selector with the BTM approach.
(c) Entropy pairs selector with the BTM approach.
(d) Random pairs selector with the Bayes approach.
(e) No repeating pairs selector with the Bayes approach.
(f) Entropy pairs selector with the Bayes approach.

Fig. 8. An illustration of the statistical comparison of results of the random (8a, 8d), no repeating pairs (8b, 8e) and entropy (8c, 8f) 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^E showing the best median performance in 18 out of the 20 distinct experiments.

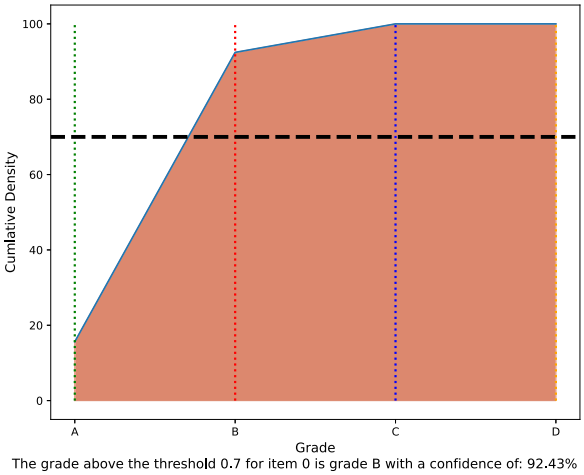


(a) Random pairs selector JSD median Results.
(b) No repeating pairs selector JSD median Results.
(c) Entropy pairs selector JSD median Results.

Fig. 9. 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*).



(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.

Fig. 10. 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 10b the panel on the *right* shows the likely grade that meets the threshold level set by the user.

Ns, as for unknown uncertainty densities over targets, no repeating pairs may not perform as well.

Unsurprisingly, comparing Fig. 8 and 9, it is evident that BCJ is better at estimating the expected rank than the complete density of the rank distribution. For example, in Fig. 8, 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 Fig. 9, 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.

5.3. 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 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. (2022).

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:

$$P(g \leq r_i \leq h) = \sum_{k=g}^{h} P(r_i = k), \tag{20}$$

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).

Fig. 10 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 Fig. 10) 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 Fig. 3. We expect that such probabilistic reasoning renders the assessors greater control over the whole CJ process, with a high level of explainability.

6. Bayesian comparative judgement on a real comparative judgement dataset

In 2018, Bramley and Vitello (2019) used a round-robin approach for selecting pairs, much like the no repeating pairs approach used in this paper, 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

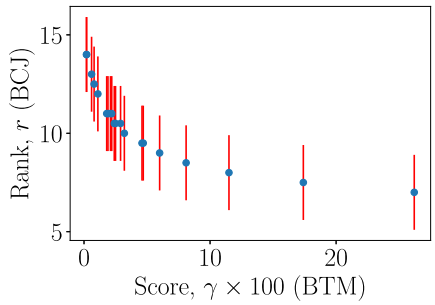


Fig. 11. Comparison between the estimated ranks r_i using BCJ and scores $100\times \gamma_i$ using BTM (see equation: (4)). The blue dots show the expected rank $\mathbb{E}[r_i]$ versus the score, and the error bar (in red) shows the standard deviation of the predicted distribution over an item's rank. The full predictive distributions are shown in Fig. 12. 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.

our demonstration with real data comparing BCJ and BTM in this section, we selected the data from study 1*b*. 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 sore was reported as 0.818, which is deemed as highly reliable. The scope and breadth of this dataset is similar to the synthetic experiments illustrated in earlier sections of this paper, 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 Fig. 11, 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 Fig. 11, 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 Fig. 12 being similar in shape and width. We 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 clear depiction of uncertainties of predictions.

In Fig. 13, 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 on 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 5.3.

7. Conclusions

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 its potential impact on various forms of assessment and associated practices (Dwivedi et al., 2023; Watermeyer et al., 2023).

With the introduction of CJ this has helped alleviate a lot of the quality issues in principle but 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 (Holmes et al., 2020). 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.

We intend in future studies to use the proposed approach to the CJ process with educators. In both quantitative and qualitative manner, we will seek to answer what works and what doesn't, and how to scale BCJ to real-world studies with potentially a large number of items while reducing the cognitive load of many assessors.

Statements on open data and ethics

In this study, we primarily used synthetic data that involves no humans, and we described how they were generated in Section 3.2.3. Thus, no ethics approval was required. However, this data can be obtained by sending request e-mails to the corresponding author of this paper.

For the real data used in Section 6, 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 of Bramley and Vitello (2019). Hence, we were merely the users of anonymised data, and since it is owned by the authors of Bramley and Vitello (2019), 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).

A code example of Bayesian Comparative Judgement is available at: https://github.com/codingWithAndy/Bayesian-Comparative-Judgement.

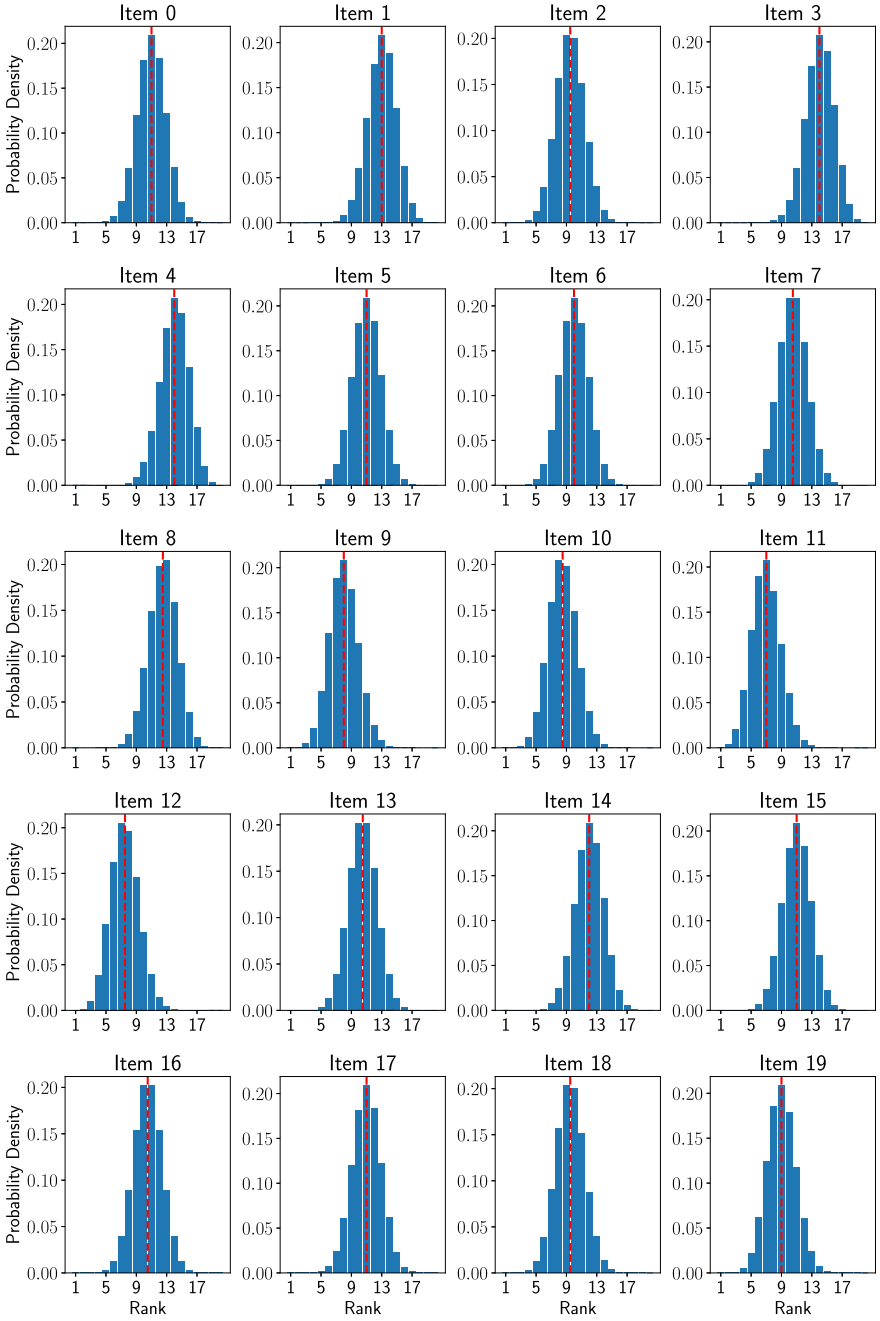


Fig. 12. Illustration of the predictive probability distribution generated using BCJ along with the $\mathbb{E}[r_i]$ for each item i (depicted with red dotted vertical lines) using dataset 1*b* from (Bramley & Vitello, 2019). The experiment had an SSR score of 0.818, which is considered a respectable level for CJ as it is above the minimum of 0.7.

CRediT authorship contribution statement

Andy Gray: Writing – review & editing, Writing – original draft, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alma Rahat:** Writing – review & editing, Writing – original draft, Conceptualization, Supervision. **Tom Crick:** Writing – original draft, Supervision. **Stephen Lindsay:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project was funded by the EPSRC Centre for Doctoral Training in *Enhancing Human Interactions and Collaborations with Data and Intelligence-Driven Systems* (EP/S021892/1) based at Swansea University. The industry partner for this co-funded PhD project is CDSM Interactive Solutions Ltd; we are particularly grateful to their Chief Technical Officer, Darren Wallace, for invigorating the discussion around the topic and guiding the application's direction. We would also like to thank Jennifer Pearson for their valuable feedback on the manuscript.

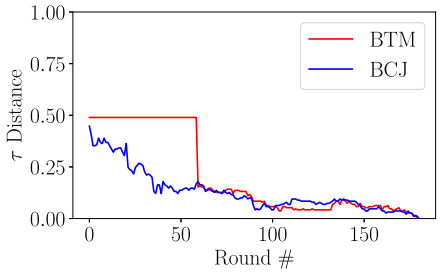


Fig. 13. A comparison between the convergence of the BTM CJ (red line) and BCJ (blue line) against their respective final ranks. The BTM approach took ≈ 60 comparisons before generating a reasonable rank. Until this point, it produced a flat τ distance value of 0.5. Our Novel BCJ approach started to 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.

the authors of Bramley and Vitello (2019) for providing us with the anonymised data from their experiment.

References

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

Finn, P., & Cinpoes, R. (2022). 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.

Nisbet, I., & Shaw, S. (2020). Sage: Is Assessment Fair?.

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

Everett, N., & Papageorgiou, J. (2011). *Investigating the accuracy of predicted: A level grades as part of 2009 UCAS admission process*. Department for Business Innovation & Skills.

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

Crick, T., Knight, C., Watermeyer, R., & Goodall, J. (2020). The impact of COVID-19 and "Emergency Remote Teaching" on the UK Computer Science Education Community, https://doi.org/10.1145/3416465.3416472.

Marchant, E., Todd, C., James, M., Crick, T., Dwyer, R., & Brophy, S. (2021). 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*, *16*(12), Article e0260396. https://doi.org/10.1371/journal.pone.0260396.

Crick, T., Knight, C., Watermeyer, R., & Goodall, J. (2021). 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)* (pp. 1048–1055).

Siegel, A., Zarb, M., Alshaigy, B., Blanchard, J., Crick, T., Glassey, R., Holt, J. R., Latulipe, C., Riedesel, C., Senapathi, M., Simon, & Williams, D. (2021). Teaching through a global pandemic: Educational landscapes before, during and after COVID-19. In *Proceedings of the 2021 working group reports on innovation and technology in computer science education (ITiCSE-WGR'21)*.

Lowthian, E., Abbasizanjani, H., Bedston, S., Akbari, A., Cowley, L., Fry, R., Owen, R. K., Hollinghurst, J., Rudan, I., Beggs, J., Marchant, E., Torabi, F., de Lusignan, S., Crick, T., Moore, G., Sheikh, A., & Lyons, R. A. (2023). 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*. https://doi.org/10.1177/01410768231181268.

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

Shankar, K., Phelan, D., Suri, V., Watermeyer, R., Knight, C., & Crick, T. (2021). "The COVID-19 Crisis is not the core problem": Experiences, challenges, and concerns of Irish academia in the pandemic. *Irish Educational Studies*, *40*(2), 169–175. https://doi.org/10.1080/03323315.2021.1932550.

McGaughey, F., Watermeyer, R., Shankar, K., Suri, V., Knight, C., Crick, T., Hardman, J., Phelan, D., & Chung, R. (2022). 'this can't be the new norm': Academics' perspectives on the COVID-19 crisis for the Australian university sector. *Higher Education Research and Development*, *41*(7). https://doi.org/10.1080/07294360.2021.1973384.

Hardman, J., Watermeyer, R., Shankar, K., Suri, V., Crick, T., Knight, C., McGaughey, F., & Chung, R. (2022). "Does anyone even notice us?" COVID-19's impact on academics' well-being in a developing country. *South African Journal of Higher Education*, *36*(1), 1–19. https://doi.org/10.20853/36-1-4844.

Crick, T. (2021). COVID-19 and digital education: A catalyst for change? *ITNOW*, *63*(1). https://doi.org/10.1093/itnow/bwab005.

Ward, R., Phillips, O., Bowers, D., Crick, T., Davenport, J. H., Hanna, P., Hayes, A., Irons, A., & Prickett, T. (2021). Towards a 21st Century personalised learning skills taxonomy. In *Proceedings of IEEE global engineering education conference (EDUCON'21)* (pp. 344–354).

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

Irons, A., & Crick, T. (2022). 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* (pp. 231–244). Emerald Publishing.

Crick, T., Knight, C., & Watermeyer, R. (2022). 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)*.

Thomas, E., Crick, T., & Beauchamp, G. (2023). Envisioning the post-COVID "new normal" for education in Wales. *Wales Journal of Education*, *25*(2). https://doi.org/10.16922/wje.25.2.1.

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

Crick, T., Prickett, T., & Bradnum, J. (2022). 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)* (pp. 519–525).

Ward, R., Crick, T., Davenport, J. H., Hanna, P., Hayes, A., Irons, A., Miller, K., Moller, F., Prickett, T., & Walters, J. (2023). Using skills profiling to enable badges and micro-credentials to be incorporated into higher education courses. *Journal of Interactive Media in Education*, *2023(1)*(10), 1317–1336. https://doi.org/10.5334/jime.807.

Knight, C., Conn, C., Crick, T., & Brooks, S. (2023). Divergences in the framing of inclusive education across the UK: A four nations critical policy analysis. *Educational Review*. https://doi.org/10.1080/00131911.2023.2222235.

Weale, S. (2022). 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.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. (2016). *Intelligence unleashed: An argument for AI in education*. Tech. rep. Pearson Education.

Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, *11*(1), 237. https://doi.org/10.3390/app11010237.

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences*, *10*(3), 1042. https://doi.org/10.3390/app10031042.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kumar Kar, A., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., ... Williams, M. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *53*, Article 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002.

Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). Student retention using educational data mining and predictive analytics: A systematic literature review. *IEEE Access*. https://doi.org/10.1109/ACCESS.2022.3188767.

Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., & Rangwala, H. (2016). Predicting student performance using personalized analytics. *Computer*, *49*(4), 61–69. https://doi.org/10.1109/MC.2016.119.

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(11). https://doi.org/10.1186/s40561-022-00192-z.

Iqbal, Z., Qadir, J., Noor Mian, A., & Kamiran, F. (2017). Machine learning based student grade prediction: A case study, https://doi.org/10.48550/arXiv.1708.08744.

Vijayalakshmi, V., & Venkatachalapathy, K. (2019). Comparison of predicting student's performance using machine learning algorithms. *International Journal of Intelligent Systems Technologies and Applications*, *12*, 34–45. https://doi.org/10.5815/ijisa.2019.12.04.

Yousafzai, B., Hayat, M., & Afzal, S. (2020). Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. *Education and Information Technologies*, *25*, 4677–4697. https://doi.org/10.1007/s10639-020-10189-1.

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, *57*(10), 1510–1529. https://doi.org/10.1177/0002764213479366.

Williamson, B., Bayne, S., & Shay, S. (2020). The datafication of teaching in higher education: Critical issues and perspectives. *Teaching in Higher Education*, *25*(4), 351–365. https://doi.org/10.1080/13562517.2020.1748811.

Akgun, S., & Greenhow, C. (2019). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, *2*, 431–440. https://doi.org/10.1007/s43681-021-00096-7.

Williamson, B., Eynon, R., & Potter, J. (2020). Pandemic politics, pedagogies and practices: Digital technologies and distance education during the coronavirus emergency.

Learning, Media and Technology, *45*(2), 107–114. https://doi.org/10.1080/17439884.2020.1761641.

Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I* (pp. 99–127). World Scientific.

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

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. *Studies in Language Testing*, *3*, 74–91.

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, *32*(1), 384–406. https://doi.org/10.1214/aos/1079120141.

Coenen, T., Coertjens, L., Vlerick, P., Lesterhuis, M., Mortier, A. V., Donche, V., Ballon, P., & De Maeyer, S. (2018). An information system design theory for the comparative judgement of competences. *European Journal of Information Systems*, *27*(2), 248–261. https://doi.org/10.1080/0960085X.2018.1445461.

Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgment*. Tech. rep. Cambridge Assessment.

Holmes, S., Black, B., & Morin, C. (2020). *Marking reliability studies 2017: Rank ordering versus marking – which is more reliable?* Tech. rep. Ofqual.

Bramley, T., & Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education*, *26*(1), 43–58.

Chen, O., Paas, F., & Sweller, J. (2023). A cognitive load theory approach to defining and measuring task complexity through element interactivity. *Educational Psychology Review*, *35*(63). https://doi.org/10.1007/s10648-023-09782-w.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144. https://doi.org/10.1007/BF00117714.

Bramley, T. (2007). Paired comparison methods. In *Techniques for monitoring the comparability of examination standards* (pp. 246–300).

Bartholomew, S. R., Strimel, G. J., & Yoshikawa, E. (2019). Using adaptive comparative judgment for student formative feedback and learning during a middle school design project. *International Journal of Technology and Design Education*, *29*(2), 363–385. https://doi.org/10.1007/s10798-018-9442-7.

Christodoulou, D. (2017). *Making good progress?: The future of assessment for learning*. Oxford University Press.

Pollitt, A. (2004). Let's stop marking exams. In *IAEA conference*. University of Cambridge Local Examinations Syndicate.

Pinot de Moira, A., Wheadon, C., & Christodoulou, D. (2022). The classification accuracy and consistency of comparative judgement of writing compared to rubric-based teacher assessment. *Research in Education*, *113*(1), 25–40. https://doi.org/10.1177/00345237221118116.

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

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). Scale separation reliability: What does it mean in the context of comparative judgment? *Applied Psychological Measurement*, *42*(6), 428–445. https://doi.org/10.1177/0146621617748321.

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2020). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education*, *27*(1), 46–64. https://doi.org/10.1080/0969594X.2019.1700212.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, *39*(3–4), 324–345. https://doi.org/10.1093/biomet/39.3-4.324.

Luce, R. D. (1959). *Individual choice behavior*.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. https://doi.org/10.1007/BF02293814.

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*, *29*(3), 211–223. https://doi.org/10.1080/08957347.2016.1171769.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2002). *Applied statistics for the behavioural sciences* (6th edition). Houghton Mifflin.

Jones, I., & Davies, B. (2022). Comparative judgement in education research. *International Journal of Research and Method in Education*. https://doi.org/10.1080/1743727X.2023.2242273.

Kelly, K. T., Richardson, M., & Isaacs, T. (2022). Critiquing the rationales for using comparative judgement: A call for clarity. *Assessment in Education*, *29*(6), 674–688. https://doi.org/10.1080/0969594X.2022.2147901.

Bloxham, S., & Price, M. (2015). External examining: Fit for purpose? *Studies in Higher Education*, *40*(2), 195–211.

O'Connell, B., De Lange, P., Freeman, M., Hancock, P., Abraham, A., Howieson, B., & Watty, K. (2016). Does calibration reduce variability in the assessment of accounting learning outcomes? *Assessment and Evaluation in Higher Education*, *41*(3), 331–349.

Schoepp, K., Danaher, M., & Ater Kranov, A. (2019). An effective rubric norming process. *Practical Assessment, Research and Evaluation*, *23*(1), 11.

Wammes, D., Slof, B., Schot, W., & Kester, L. (2022). Pupils' prior knowledge about technological systems: Design and validation of a diagnostic tool for primary school teachers. *International Journal of Technology and Design Education*, *32*(5), 2577–2609.

Leech, T., & Chambers, L. (2022). How do judges in comparative judgement exercises make their judgements? *Research Matters*, *33*, 31–47.

Elander, J., & Hardman, D. (2002). An application of judgment analysis to examination marking in psychology. *British Journal of Psychology*, *93*(3), 303–328.

Bisson, M.-J., Gilmore, C., Inglis, M., & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education*, *2*(2), 141–164. https://doi.org/10.1007/s40753-016-0024-3.

Marshall, N., Shaw, K., Hunter, J., & Jones, I. (2020). Assessment by comparative judgement: An application to secondary statistics and English in New Zealand. *New Zealand Journal of Educational Studies*, *55*, 49–71. https://doi.org/10.1007/s40841-020-00163-3.

Gray, A., Rahat, A. A., Crick, T., Lindsay, S., & Wallace, D. (2022). Using Elo rating as a metric for comparative judgement in educational assessment. In *Proceedings of 6th international conference on education and multimedia technology (ICEMT 2022)* (pp. 272–278).

Gescheider, G. A. (2013). *Psychophysics: The fundamentals*. Psychology Press.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., et al. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1). https://doi.org/10.1038/s43586-020-00001-2.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and STAN*.

Lambert, B. (2018). A student's guide to bayesian statistics. *A Student's Guide to Bayesian Statistics*, 1–520.

Pritikin, J. N. (2020). An exploratory factor model for ordinal paired comparison indicators. *Heliyon*, *6*(9), Article e04821. https://doi.org/10.1016/j.heliyon.2020.e04821.

Wainer, J. (2022). A Bayesian Bradley-Terry model to compare multiple ML algorithms on multiple data sets, https://doi.org/10.48550/arXiv.2208.04935.

Tsukida, K., & Gupta, M. R. (2011). *How to analyze paired comparison data*. Tech. rep. UWEETR-2011-0004 Department of Electrical Engineering, University of Washington.

De Maeyer, S. (2021). Bayesian analysis of comparative judgement data. https://svendemaeyer.netlify.app/posts/2021-01-18-bayesian-analysis-of-comparative-judgement-data/.

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

Fink, D. (1997). *A compendium of conjugate priors*. Tech. rep.

Feller, W. (1968). Stirling's formula. In *An introduction to probability theory and its applications* (pp. 50–53). Wiley.

Mackay, D. J. C. (1998). Introduction to Monte Carlo methods. In *Learning in graphical models* (pp. 175–204). Springer.

Koehler, E., Brown, E., & Haneuse, S. J.-P. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, *63*(2), 155–162. https://doi.org/10.1198/tast.2009.0030.

Hughes, E. J. (2001). Evolutionary multi-objective ranking with uncertainty and noise. In *LNCS: Vol. 1993. International conference on evolutionary multi-criterion optimization (EMO 2001)* (pp. 329–343).

Andrews, L. C. (1998). *Special functions of mathematics for engineers*. Oxford University Press.

Thiagarajan, P., & Ghosh, S. (2023). Jensen-Shannon divergence based novel loss functions for Bayesian neural networks, https://doi.org/10.48550/arXiv.2209.11366.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2), 281–305.

Settles, B. (2010). *Active learning literature survey*. Tech. rep., Computer sciences technical report 1648 University of Wisconsin-Madison.

Knijnenburg, B. P., & Willemsen, M. C. (2015). Evaluating recommender systems with user experiments. In *Recommender systems handbook* (pp. 309–352). Springer.

Das, S., Wong, W.-K., Dietterich, T., Fern, A., & Emmott, A. (2016). Incorporating expert feedback into active anomaly discovery. In *IEEE 16th international conference on data mining (ICDM 2016)* (pp. 853–858).

MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation*, *4*(4), 590–604. https://doi.org/10.1162/neco.1992.4.4.590.

Zhan, X., Liu, H., Li, Q., & Chan, A. B. (2021). A comparative survey: Benchmarking for pool-based active learning. In *Proceedings of 30th international joint conference on artificial intelligence (IJCAI-21)* (pp. 4679–4686).

Lewis, D. D. (1995). A sequential algorithm for training text classifiers: Corrigendum and additional data. *ACM SIGIR Forum*, *29*(2), 13–19. https://doi.org/10.1145/219587.219592.

Lazo, A. V., & Rathie, P. (1978). On the entropy of continuous probability distributions (corresp.). *IEEE Transactions on Information Theory*, *24*(1), 120–122. https://doi.org/10.1109/TIT.1978.1055832.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... van Mulbregt, P. (2020). SciPy 1.0 contributors, SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Miller, R. G. (1981). *Simultaneous statistical inference*.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1–2), 81–93. https://doi.org/10.1093/biomet/30.1-2.81.

Fagin, R., Kumar, R., & Sivakumar, D. (2003). Comparing top k lists. *SIAM Journal on Discrete Mathematics*, *17*(1), 134–160. https://doi.org/10.1137/S0895480102412856.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E., Jeyaraj, A., Kar, A., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Al-Bashrawi, M., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L., … Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, Article 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642.

Watermeyer, R., Phipps, L., Lanclos, D., & Knight, C. (2023). *Generative AI and the automating of academia*. Postdigital Science and Education.