



Gray, A. (2026) Bayesian comparative judgement in the wild. *Royal Statistical Society International Conference*, Bournemouth, UK, 7-10 September 2026.

ResearchSPAce

<https://researchspace.bathspa.ac.uk/>

Your access and use of this document is based on your acceptance of the ResearchSPAce Metadata and Data Policies, as well as applicable law:- <https://researchspace.bathspa.ac.uk/policies.html>

Unless you accept the terms of these Policies in full, you do not have permission to download this document.

This cover sheet may not be removed from the document.

Please scroll down to view the document.

Bayesian Comparative Judgement in the Wild

Dr Andy Gray

Contributors: Alma Rahat, Tom Crick, Stephen Lindsay

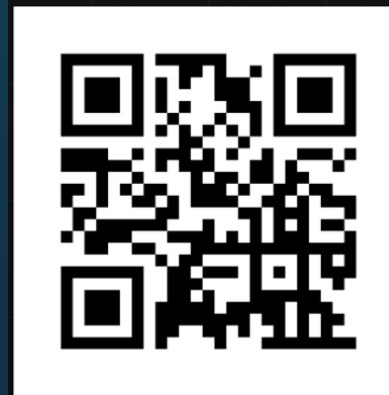
ROYAL STATISTICAL SOCIETY CONFERENCE 2026 · BOURNEMOUTH

About Me

- Senior Lecturer in Computing at Bath Spa University
- X: @codingWithAndy
- ORCID: <https://orcid.org/0000-0002-1150-2052>
- Email: a.gray2@bathspa.ac.uk

Research Papers

1. A Bayesian active learning approach to comparative judgement within education assessment
2. Bayesian Active Learning for Comparative Judgement: Estimating Reliability and Managing Multiple Criteria
3. Rendering Transparency to Ranking in Educational Assessment via Bayesian Comparative Judgement

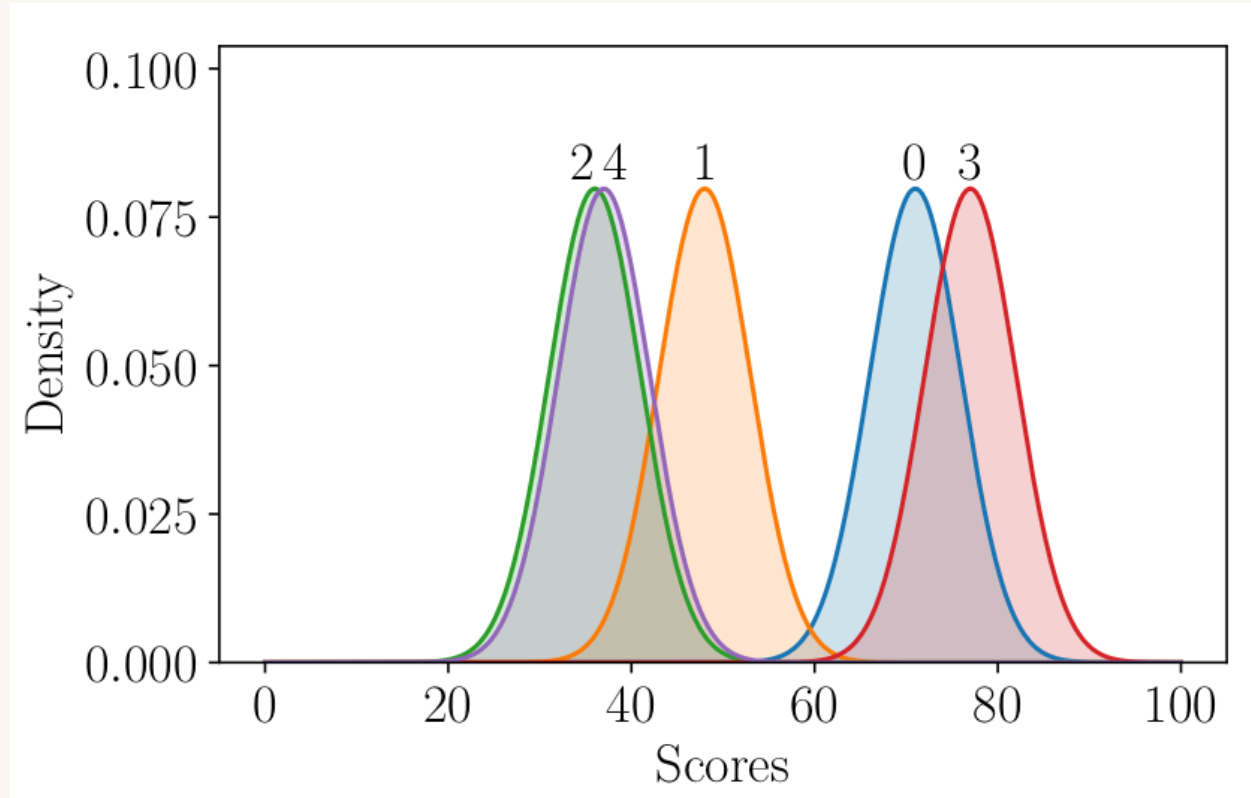


Session Overview

1. Bayesian CJ Performance
2. BCJ vs Multi-Criterion BCJ
3. Real World Experiment

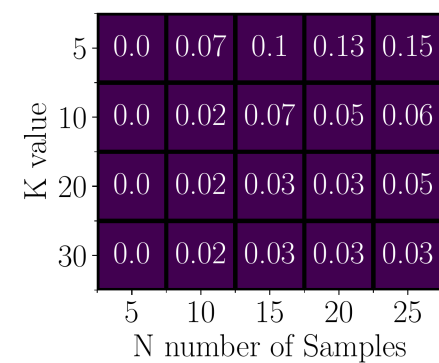
Bayesian CJ Performance

Sampling Distribution Example

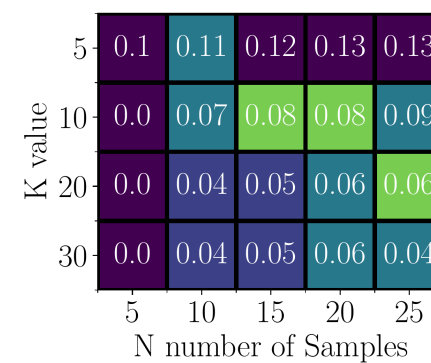


BCJ vs Results

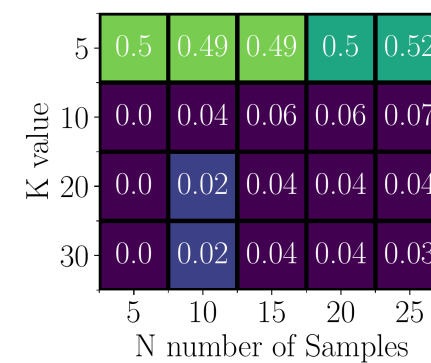
Entropy



Random Pairs

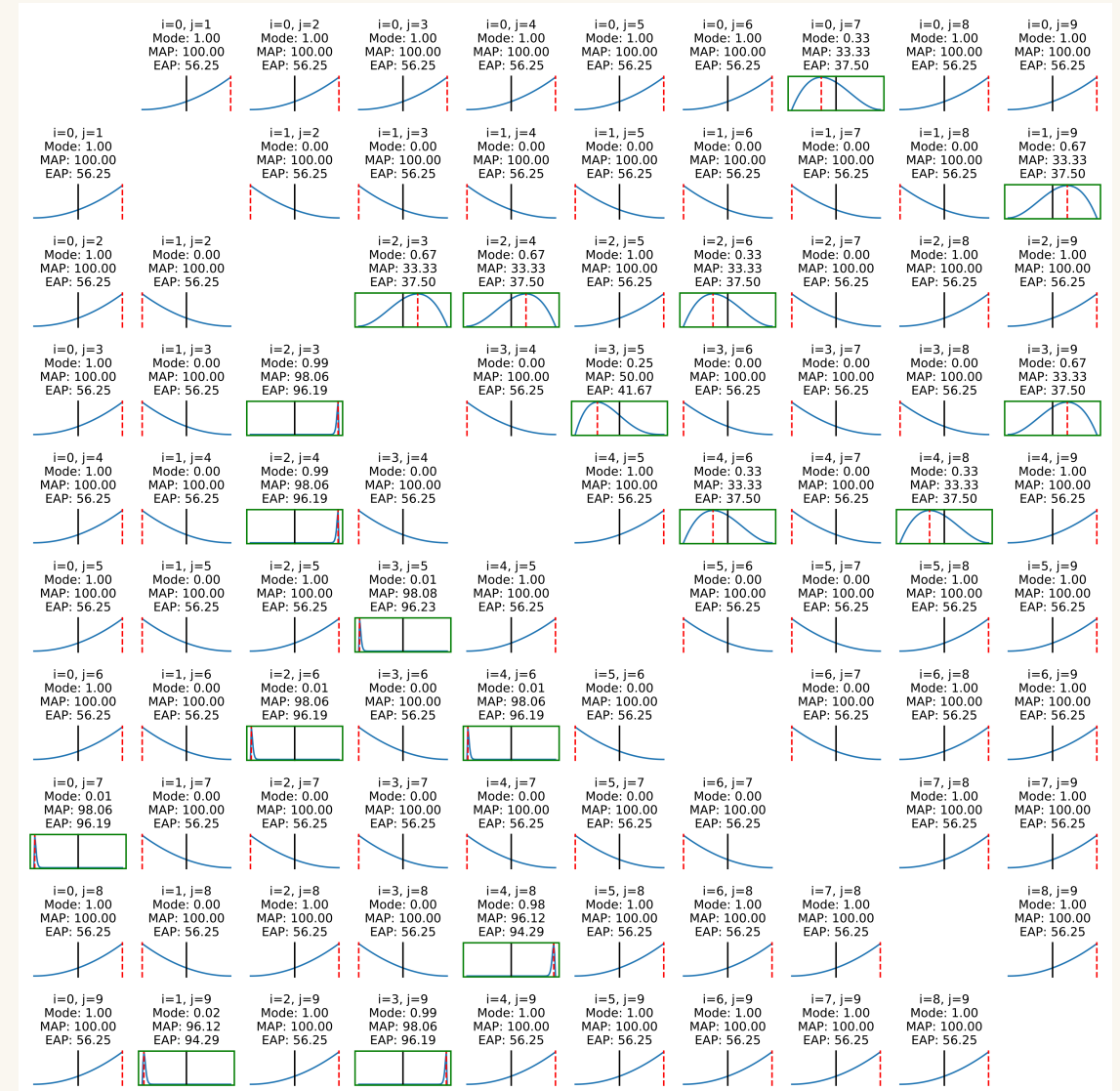


No Repeating Pairs



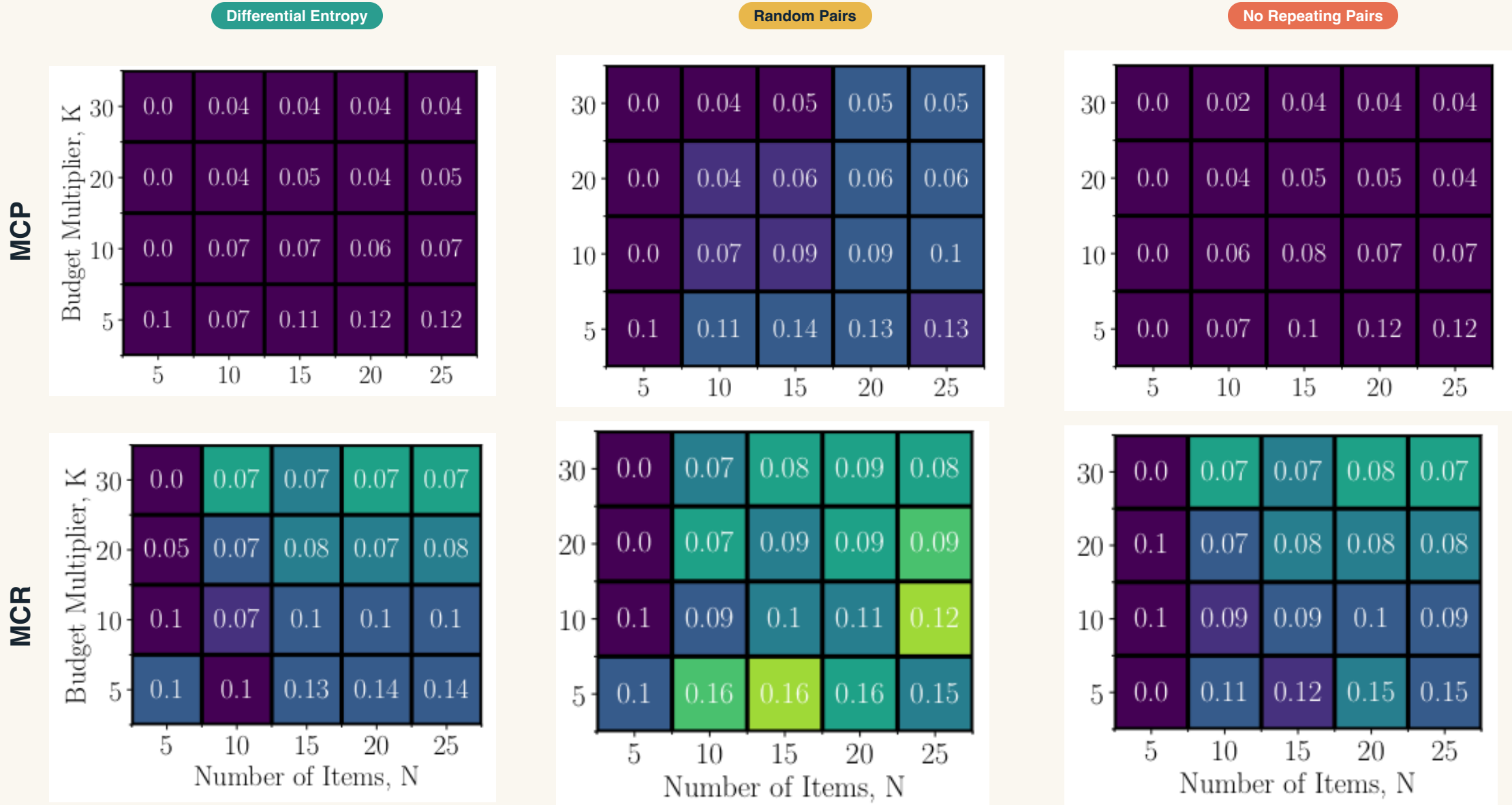
MAP & EAP

- Pre-Intervention: Kendal τ Score = 0.07
- Post-Intervention: Kendal τ Score = 0.04

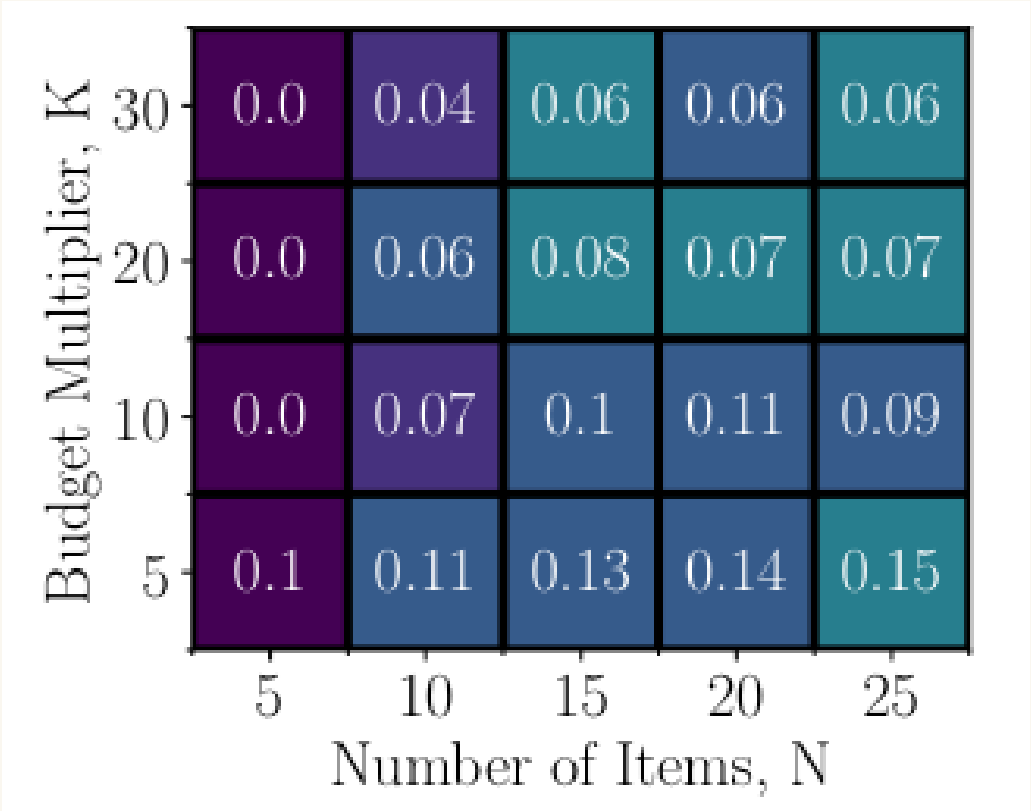


Multi-Criteria Bayesian Comparative Judgement

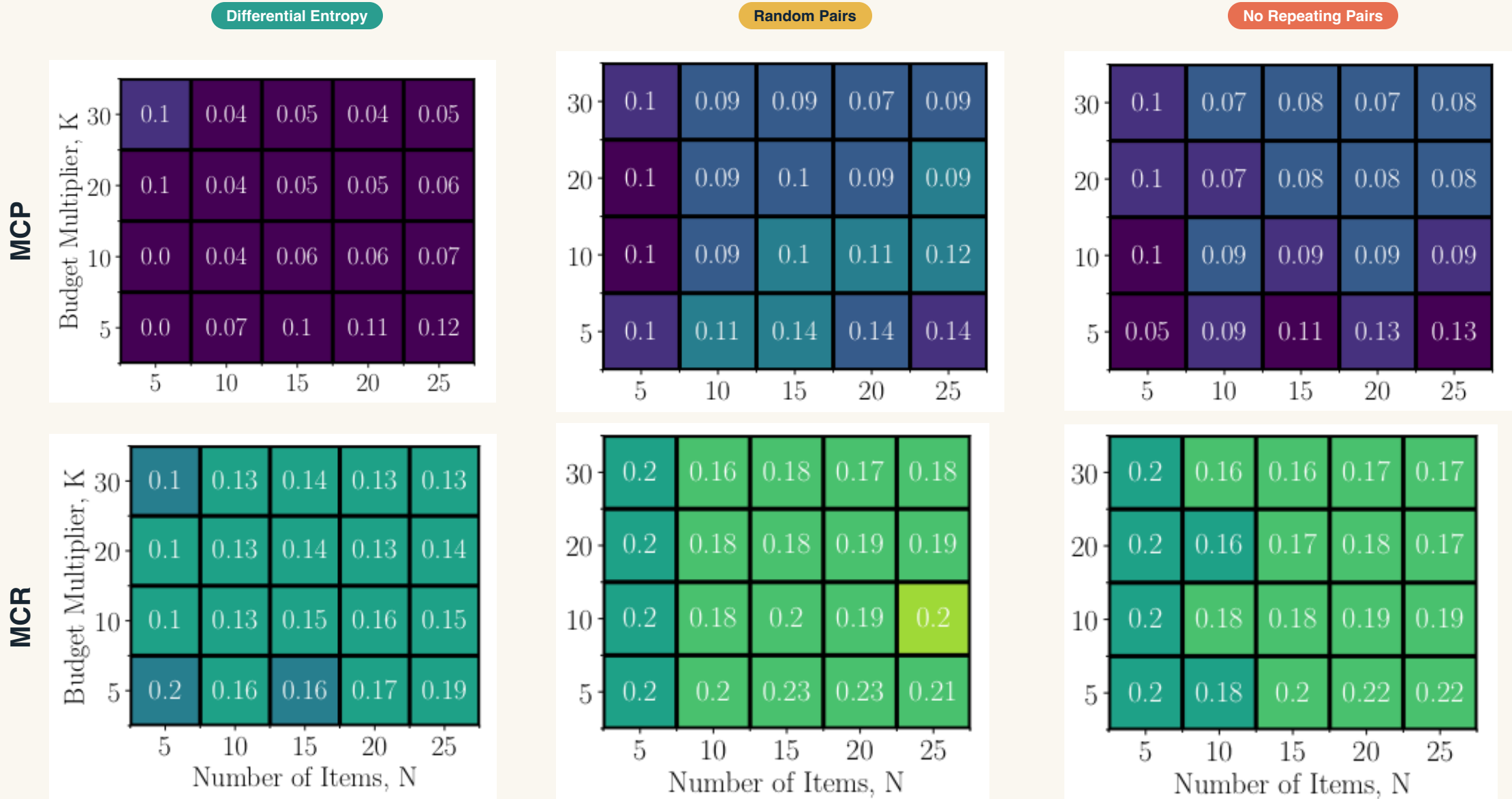
Multi-Criteria BCJ vs BCJ - DREsS Dataset



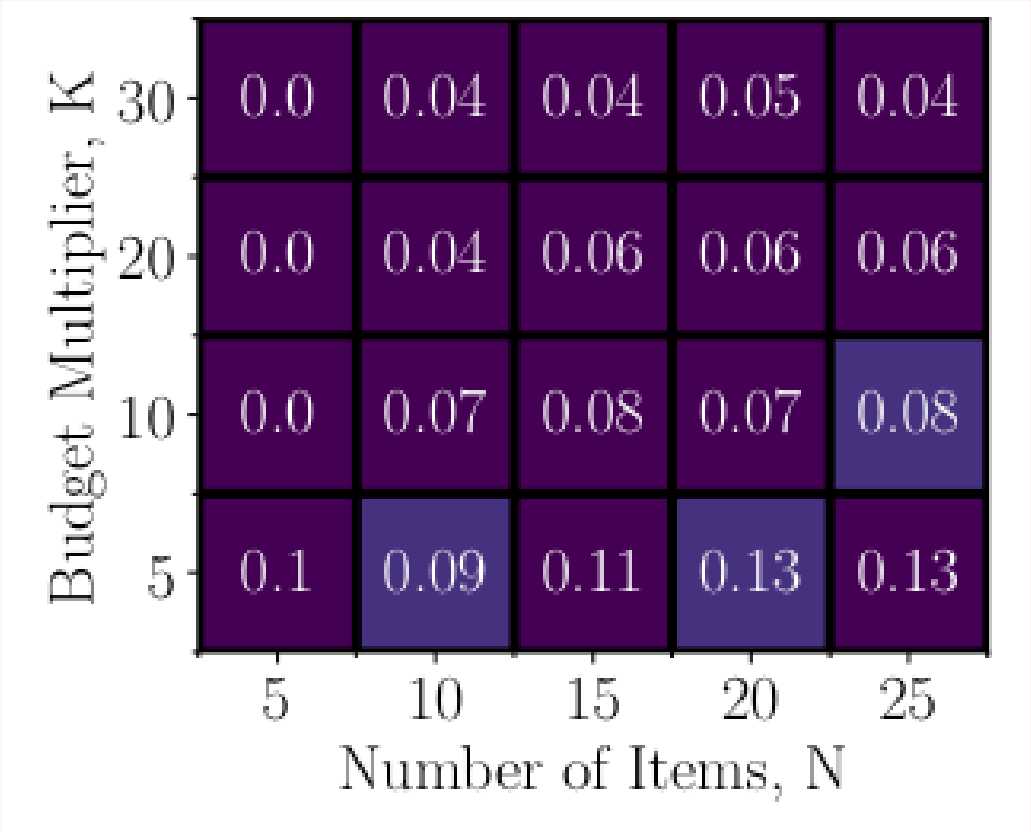
Multi-Criteria BCJ vs BCJ - DREsS Dataset



Multi-Criteria BCJ vs BCJ - BSU Dataset



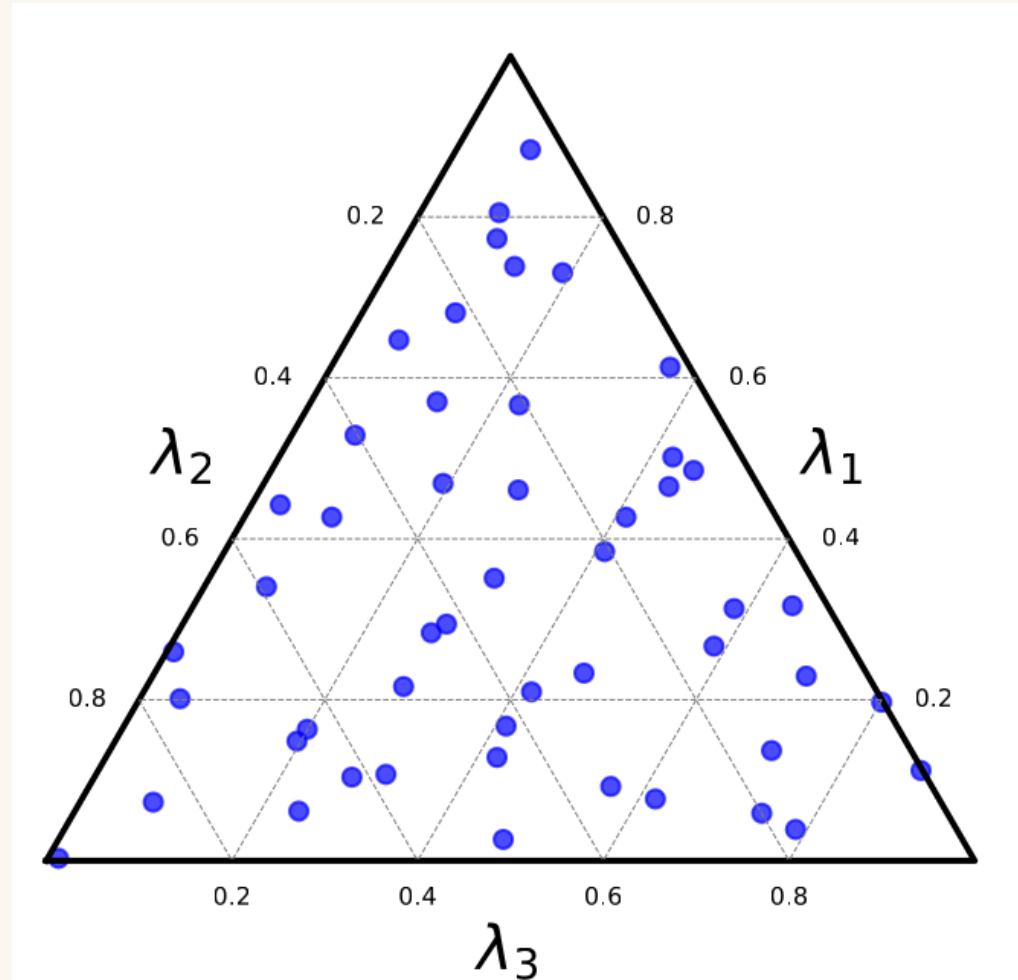
Multi-Criteria BCJ vs BCJ - BSU Dataset



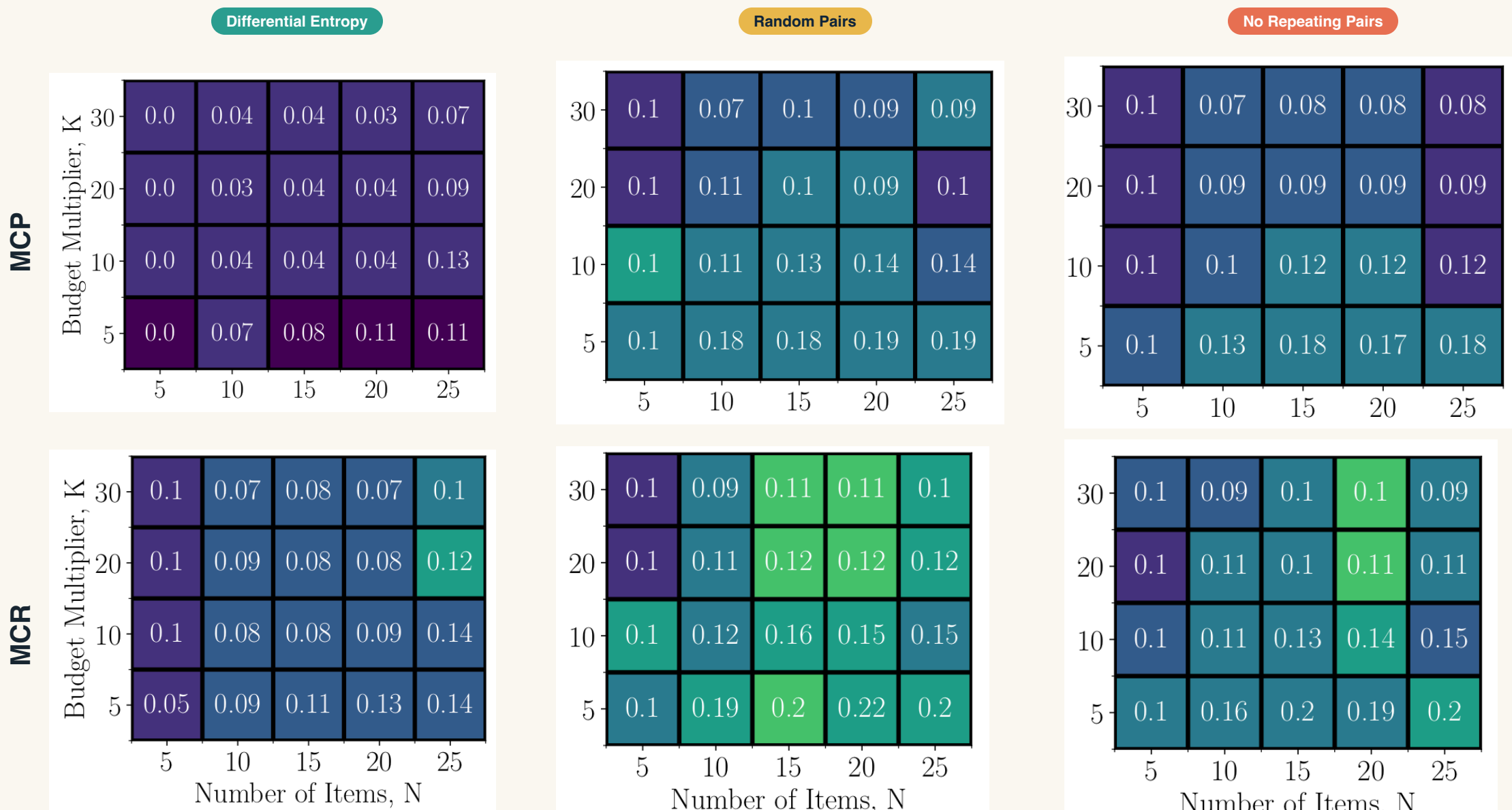
Multi-Criteria BCJ vs BCJ

WITH RANDOM WEIGHTS

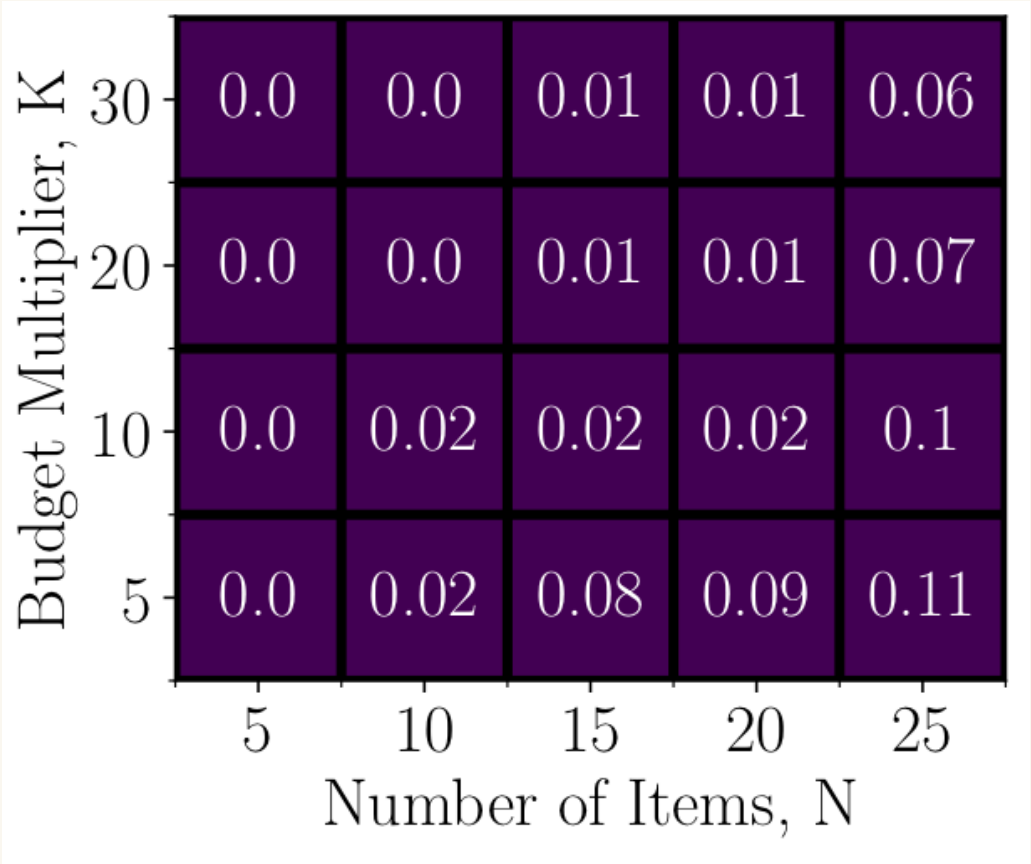
QMC Weights Transformed onto a Simplex



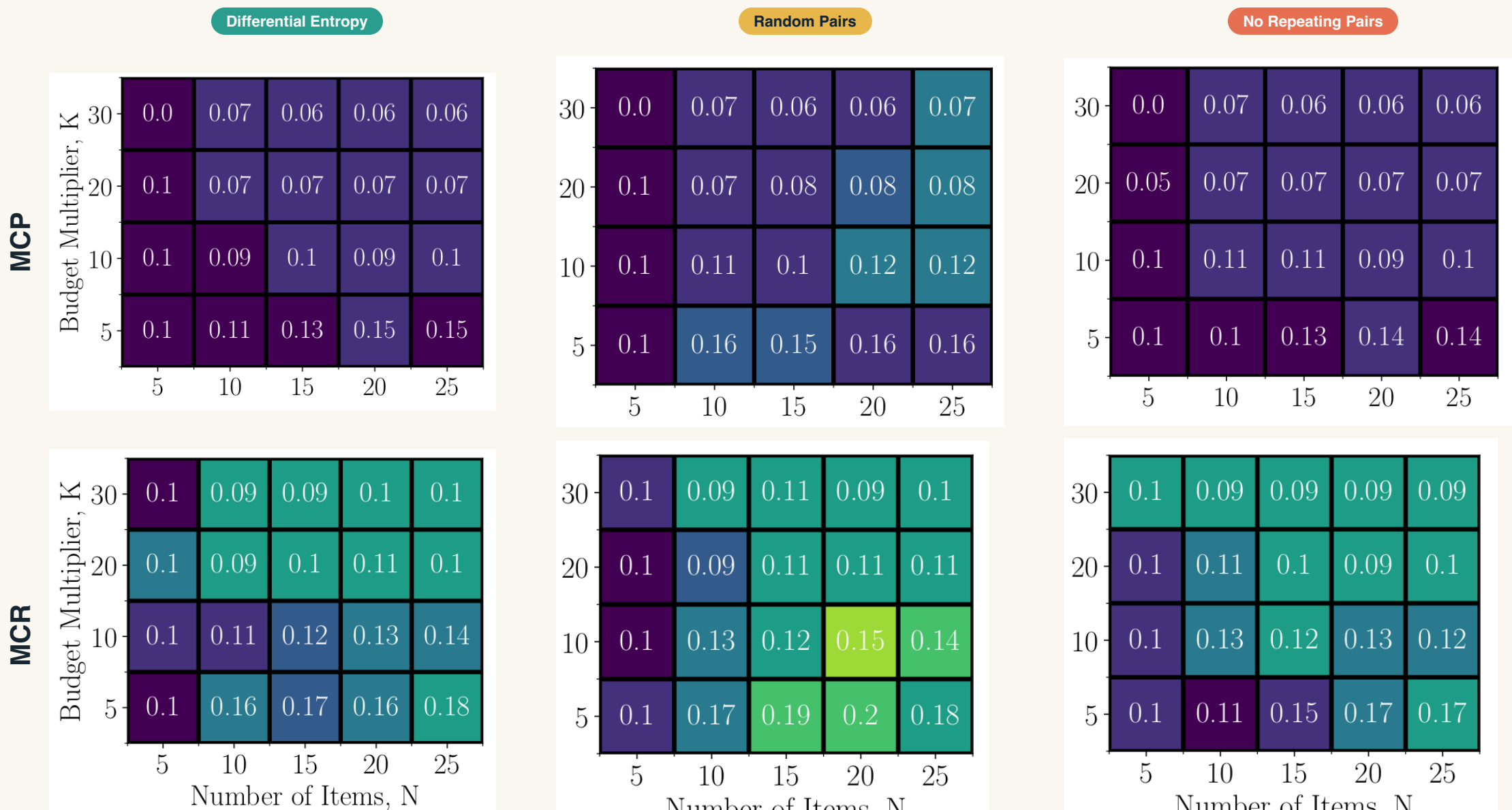
Multi-Criteria BCJ vs BCJ - DREsS Dataset



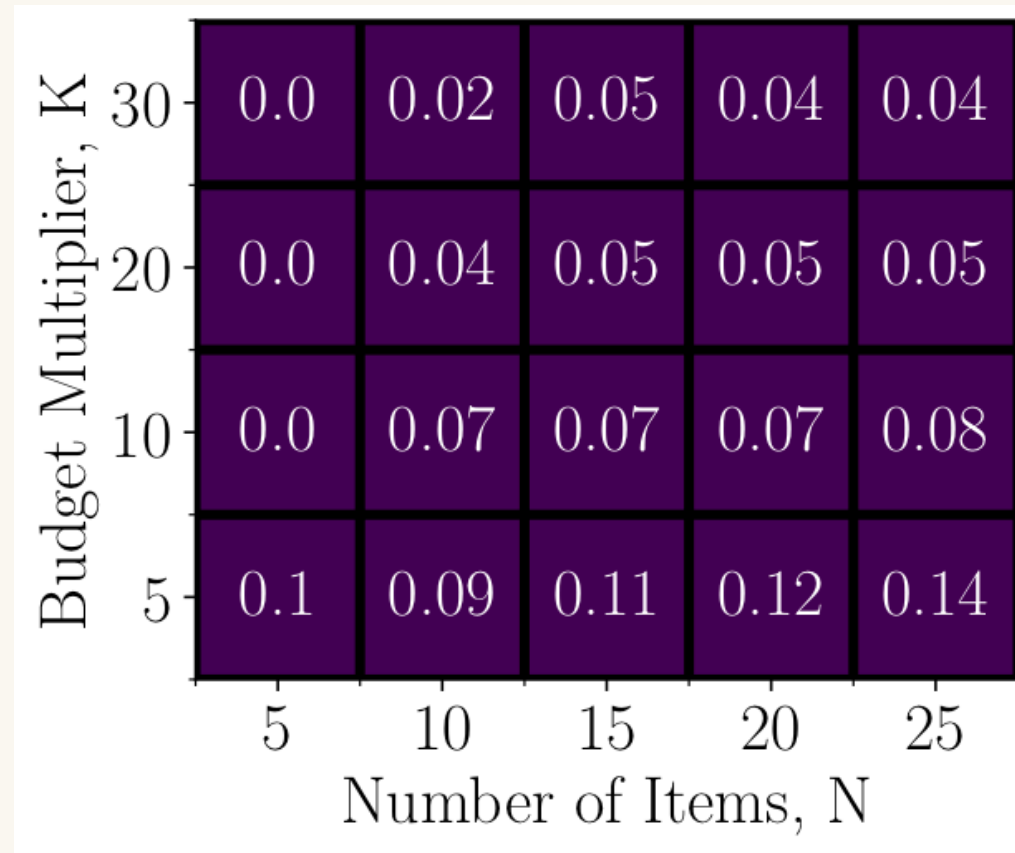
Multi-Criteria BCJ vs BCJ - DREsS Dataset



Multi-Criteria BCJ vs BCJ - BSU Dataset



Multi-Criteria BCJ vs BCJ - BSU Dataset



WHEN IN THE HEANDS OF REAL HUMANS

Real World BCJ Study

Real Markers, Real Stakes

- **30 MSc critical-review submissions**, split into 3 matched groups of 10.
- **3 experienced markers**, each completing all three methods:
 - Traditional (TAM),
 - BCJ,
 - MBCJ
- Benchmarked against the **Oracle**: the marks originally awarded and released to students.
- CJ methods capped at **2 hours**; traditional marking left open-ended, as it must run to completion.

LETS TAKE A LOOK

BCJ Web App

Word Count: 1040

CRITICALLY REVIEWING A PAPER

1

Table of Contents

Introduction.....3

Summary.....3

Evaluation.....3

Item 1

Critical Review

Evolutionary Minimization of Traffic Congestion

Word Count - 955

Introduction

To overcome the traffic congestion problem, research has taken into account the overall time taken for a driver to travel from point A to point B, the psychological factors of the driver, the road capacity, and suggesting routes with bounded rationality. However, usually suggesting a similar route to all drivers leads to congestion, thereby not serving a quick travel time. By proposing multiple alternatives for the same route, with some covering a longer distance, the paper aims to reduce the overall travel time for an NP-hard problem.

The Multiple-Routes Problem

When considering the roads in a city as edges and source and the destination as the targets, the multiple routes problem aims to give "n" number of routes such that unilaterally changing the route by a single driver will not lead to them reaching their destination quicker. A state of n user-restricted equilibrium is proposed. User-equilibrium can be defined as the state where there is no gain to be obtained by a single user changing their strategy alone since this solution proposes n routes. The goal is to reduce the overall travel time with all drivers allocated one of the given n routes.

Multiple Routes Evolutionary Algorithm

Here the MREA algorithm employs the concepts of genetic algorithm such as crossover, mutation and the population size to suggest n routes from the source to the target. The initial population created generates offspring via crossover operation or randomly mutating individuals. The algorithm performed well on each of the three crossover operators, and it was also found to perform better by using mutation and having a larger population size. The algorithm uses a randomized version of Dijkstra's Shortest path algorithm called the RandDijkstra (RD).

Critical Review of the Mutation Operator

A Poisson distribution determines the number of mutations to execute. Multiple mutation operators are combined except for the ExSegment mutation, as this is an expensive operation. For example, in the NewRoute mutation, it is mentioned that a route is selected randomly and replaced by the one computed by RD. This operation appears redundant as RD itself generates a random shortest path. Hence this method can be avoided as it may not add much value to the mutation operation.

For the RandomP mutation operator, a sub-segment of the route is replaced using a route randomly generated by RD. This seems redundant as RD also generates the initial route. Hence replacing a random route generated via a method with another random route generated by the same method

Item 6

may not add much value to this operation. The LinkWP operator improves the RandomP as it aims to

LETS TAKE A LOOK

MBCJ Web App

Critically Reviewing a Paper

Maximilian Böther et al. "Evolutionary minimization of traffic congestion". In: Proceedings of the Genetic and Evolutionary Computation Conference. 2021, pp. 937–945

Introduction

This paper aims to implement a Multiple-Routes Evolutionary Algorithm (MREA) to minimize traffic congestion amongst multiple vehicles. The focus of this paper is to lower the travel time of all vehicles in a system by utilizing an MREA that can achieve an at least 99.69% optimal solution and requires 40% of the time to run than that of a previously known solution [1]. Some vehicles will be given a sub-optimal path to reach an optimal solution for all vehicles. This paper also aims to further assess the effectiveness of the key algorithms used for mutators and crossover functions in the proposed evolutionary algorithm.

The proposed algorithm works by first using a modified version of Dijkstra’s algorithm to create a random path between every vehicle’s starting point, and every vehicle’s destination. It then applies a crossover function for two random individuals and repeats this process a set number of times which generates offspring for each route. For every offspring created from the crossover algorithm, the MREA proceeds to perform a random number of mutations selected randomly from various mutation algorithms.

Summary

This algorithm is designed to be used as a tool whereby the driver of a vehicle may choose their desired route from a list of routes whereby no route gives them a time advantage. The paper covers four mutator functions, and three crossover operators and ranks them against each other, showing that adding a new route is the best mutator function and that the crossover function could be either Greedy Crossover or Randomized Greedy Crossover. Because such an algorithm runs in non-deterministic polynomial-time, it is important to try and reduce the time costs where possible even in cases where the score may not be perfect. Overall, I was very satisfied that the solution provided achieved the goals of the paper and was impressed with how the way that the problem was divided into its key components with each component being explained in detail.

Analysis

The paper did an excellent job of showcasing the effectiveness of its various mutation operators and crossover operators but does little to showcase the effectiveness of its “RandDijkstra” algorithm. I believe that it would be important to see the fitness values prior to the application of the Random Path mutator to see if it significantly differs from k-Dijkstra in **Table 1**. I don’t believe it’s as fair of a comparison to use k-Dijkstra while accounting for delays as it would be to use RandDijkstra which was specifically designed to counter the congestion issues that come with Dijkstra’s algorithm in the scenario. I think it would also be interesting to see the effects of different initial pathing algorithms (with the same random modification) such as the improved Bellman-Ford algorithm [3] or Moore-algorithms [4] which has been shown to have a significantly lower average runtime than Dijkstra’s algorithm when solving the shortest path problem [5].

I feel as if the use-case for this algorithm is not explained clearly enough. Whilst this paper makes it clear that it is designed to lower traffic congestion in a set area such as the example of the city of Berlin, and showcases multiple different scenarios, I do not believe it is clear what exactly this system is designed

for. It could be for use in a satellite-navigation system; however, drivers may diverge from this path

Critical Review on “EvolutionaryMinimization of Traffic Congestion”
Word count: 926

Introduction:

Traffic congestion is a serious question for travelling drivers. Traditionally, drivers are more willing to pick up the optimal (shortest or fastest) route. That sometimes leads to the amount of drivers getting stuck in the same route because their optimal routes are the same. Due to that, people try to arrange an alternative path to an existing one which is called single alternative route. An alternative route is not an optimal route, but it can be the optimal from the overall perspective, since it reduces the traffic congestion. And usually, the cost of alternative route is affordable to drivers. In 2020, Thomas Bläsius and Maximilian Böther et al. proposed an algorithm which considers road capacities and psychological models. It is a strategic route problem designed to find the best alternative route for a given route for a group of drivers, called Single Alternative Path (SAP). Nevertheless, the SAP problem is still limited to an alternative route, and a route needs to be given as input. Therefore, on basis of SAP, this article extends it by applying it to multiple route problem and adapting mutation and crossover for finding better solutions, which results in multiple route evolutionary algorithm (MREA). It takes a start point and destination as input, and it enables drivers to distribute on various routes and minimizes the overall time.

Summary:

The purpose of this paper is to minimize total travelling time of all drivers while facing high flow of traffic. MREA is proposed to solve this problem. It applies multiple route (MR) problem extended from SAP and combines MR with mutation and crossover operators. These two random search methods are quite useful for exploring and optimizing new routes. In their experiment result, it indicates MREA has the capability of enhancing overall travelling time of the system by factors between 1.8 and 3. And MREA can achieve better performance by increasing the number of mutation operators and the size of population yields. Although it set to the minimum of mutation, crossover, and population size, it addresses the achievement of improving overall travelling time of the system by factors between 1.5 and 2.8. They also made comparison between MREA and SAP proposed by Bläsius et al. [1]. MREA with the best configuration achieves 99.69% solution quality. However, it only spends 40% of the run time. All of results proves that MREA is a great algorithm for solving MR problem and fasten overall travelling time. Last but not least, there is an interesting feature in the system. In order to account limited rationality and preferences of drivers, the system keeps balance among all routes, which all drivers are not able to enhance route condition by changing routes.

Evaluation:

Obviously, Multiple Route Evolutionary Algorithm is a powerful technique in term of saving overall travelling time. This method takes advantage of random search techniques, especially genetic algorithm. The 4 mutation operators work on varying single solution, and the crossover operator performs on combination of different solutions and generate better and unique results. In addition, the thought of applying Single Alternative Path to Multiple Route problem is brilliant. The single alternative path algorithm produces excellent results on single alternative route problem. However, single alternative route problem is a special case of MR problem. It cannot transfer the knowledge from SAP to MR problem. Therefore, it only works on single alternative route problem. MREA may perform less effectively than SAP in that special case, but it works extremely better on the other aspects.

Full results: per-marker breakdown

	TAM		BCJ						MBCJ			
Marker	Time	τ	Time	#Comp	BTM τ	BCJ τ	SSR	RSR	Time	#Comp	τ	RSR
1	2:17	0.36	1:07	49	0.24	0.16	0.67	0.07	2:00	37	0.13	-1.68
2	1:26	0.40	2:00	46	0.20	0.18	0.67	0.06	1:49	57	0.16	-0.97
3	1:54	0.40	1:22	50	0.18	0.18	0.64	0.05	1:22	53	0.29	-0.77
Combined	5:37	0.49	4:29	145	0.04	0.02	0.58	0.60	5:11	147	0.20	0.35

Combined Accuracy against the Oracle

0.49

Traditional
marking

0.02

BCJ

0.20

MBCJ

Time and effort: no accuracy tax

TRADITIONAL

112 ± 25 min

all 10 scripts marked in full

BCJ

89 ± 29 min

~48 comparisons — full
pairwise coverage

MBCJ

104 ± 20 min

~47 comparisons — near-
full coverage, per criterion

A reliability metric reality check

- The field's go-to metric, **Scale Separation Reliability (SSR)**, *fell* from ~ 0.65 (individual markers) to **0.58** once all markers' judgements were pooled
- **RSR**, by contrast, *rose* from ~ 0.06 individually to **0.60** combined

BEYOND THE NUMBERS

What did the Markers Think?

"[BCJ] felt somewhat like a black box... apart from the transparency the system produces in displaying the ranking probabilities."

— Marker, on standard BCJ

"[MBCJ] made it more transparent... I could see how individual components contributed to the overall score."

— Marker, on Multi-Criteria BCJ

Flipping Opinions

- Going in, every marker **expected traditional marking to be the most trustworthy** — it's familiar, and it comes with an explicit justification for every mark.
- Shown the τ scores side by side, the room's reaction was genuine surprise: traditional marking had the **worst** agreement with the Oracle *and* the most disagreement between markers.
- By the end, the consensus had shifted: **MBCJ** was seen as the fairer, more scalable option for multi-marker moderation — with one persistent caveat.

"The main limitation remained the lack of direct feedback... before it could fully replace conventional methods."

What the experts think: two very different vantage points

THOSE WHO DEVELOP CJ METHODS

An academic CJ researcher, using it in their own teaching since 2009: reliability comes from *volume of judgements*, not the statistics dressing it up.

THOSE WHO DEPLOY IT OPERATIONALLY

Exam-board and government-side practitioners: cautious about priors introducing bias in *high-stakes* settings, and about CJ's holistic format hiding the reasoning behind a decision, whoever is doing the ranking.

"Breaking down judgements by learning outcome makes the process feel more familiar and intuitive to educators."

— CJ expert, on MBCJ

"MBCJ more closely mirrors traditional assessment logic, where individual attributes... are considered independently."

— CJ expert, on MBCJ

Takeaways

1. **BCJ process was both faster and more accurate** than the traditional marking.
2. **MBCJ is the one markers preferred**
3. **Reliability metrics provide granular pair-wise insights**

Thank you

Research Papers

1. A Bayesian active learning approach to comparative judgement within education assessment
2. Bayesian Active Learning for Comparative Judgement: Estimating Reliability and Managing Multiple Criteria
3. Rendering Transparency to Ranking in Educational Assessment via Bayesian Comparative Judgement

