Automatic Matchmaking in Two-Versus-Two Sports

Sören Rüttgers
Faculty of Technology
Bielefeld University
sruettgers@techfak.uni-bielefeld.de

Ulrike Kuhl
Faculty of Technology
Bielefeld University
ukuhl@techfak.uni-bielefeld.de

Benjamin Paaßen
Faculty of Technology
Bielefeld University
bpaassen@techfak.uni-bielefeld.de

ABSTRACT

To train two-versus-two sports, it is beneficial to play regularly with varying teammates and opponents of similar skill level. However, even in small classes, it is almost impossible for a human instructor to maintain an accurate overview of each student's skill development to optimize teams and pairings accordingly. Therefore, we propose an educational data mining approach to automated matchmaking. In particular, we trace all players' skill levels via the glicko2 algorithm and use the resulting skill ratings to optimize matchmaking in an integer linear programming approach. We explain the resulting matches in terms of ratings and counterfactual explanations to enhance the transparency of the system for instructors and players. In addition to the algorithm, we provide an evaluation on synthetic data and in a field study ($N = 38$) conducted in a course for the fast-paced two-versus-two sport Roundnet. Our analyses show that our proposed approach outperforms all baselines in terms of minimizing skill gaps and ensuring variability among teammates and opponents. The subsequent field study corroborated the positive algorithmic evaluation by comparing the experience of participants subjected either to our proposed matchmaking approach or to a random baseline. Participants' responses indicate that our approach was perceived as more trustworthy, and the explanations associated with it were deemed to be more actionable, useful, and of higher quality.

Keywords

Matchmaking, Glicko, Roundnet, sports education, explainability

1. INTRODUCTION

Two-versus-two (2v2) sports, like doubles tennis, are popular with millions of registered players in Germany alone [10]. Players typically meet in regular training sessions, where an instructor distributes them into teams and pairs them up against other teams [33]. However, team composition and pair assignment (matchmaking, for short) is challenging: how does one ensure that each player is paired with the optimal teammates and opponents to improve their skill?

We phrase the matchmaking problem as a combination of two educational data mining problems: knowledge tracing, i.e. estimating the skill development of each player over time [1], and optimal group composition based on the skill estimates. Most prior approaches address 1v1 sports. In such cases, knowledge tracing can be performed via rating systems like ELO or glicko [12], and opponents of similar skill can be matched via simple pairwise matching algorithms. However, in a 2v2 sports setting, multiple objectives become important: a) we want to minimize the skill gap between the strongest and weakest players in a match such that all players can meaningfully contribute to the game and thus have an opportunity to learn; b) we want opposing teams to have similar skill levels; and c) we want to ensure that teammates and opponents vary to prevent over-specialization. To our knowledge, no prior matchmaking scheme has addressed this multi-objective optimization for 2v2 sports.

We provide a novel, multi-objective optimization scheme for 2v2 sports. Specifically, our contributions are as follows:

1. We formalize the 2v2 matchmaking problem and prove that 2v2 matchmaking is NP-hard.
2. We propose a heuristic by rewriting 2v2 matchmaking as an integer linear program (ILP), for which efficient heuristics exist (at least for usual class sizes of up to 40 players).
3. We implement an integrated system that traces skill development via glicko2, performs matchmaking via our proposed approach, and provides explanations for the matchmaking in terms of ratings and counterfactual explanations. The source code can be found at `https://github.com/sruettgers/automatic_matchmaking`.
4. We present two validations of our proposed approach, one on synthetic data, as well as one in a field study conducted within a Roundnet course with $N = 38$ participants, verifying that our proposed scheme outperforms baselines in terms of objective and subjective match quality as well as explainability. The study data can be found at `https://github.com/sruettgers/automatic_matchmaking`.

S. Rüttgers, U. Kuhl, and B. Paaßen. Automatic matchmaking in two-versus-two sports. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 458–468, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729860

2. BACKGROUND AND RELATED WORK

Data mining approaches are prevalent in various aspects of sports to support the decisions of trainers and managers. Examples include talent identification [27, 36, 18], evaluating and improving training regimens [6, 17, 39], assessing the potential for injury [31, 29, 37], and predicting competitive performance [28, 34, 7]. More related to our work, some prior approaches also used machine learning and statistical-based approaches to aid managerial decisions by evaluating potential candidate contributions to a team's forecasted winning percentage [43], estimating team performance and predicting transfer outcomes [16], and exploring the impact of match history on predicting athletic success [23].

While these approaches offer promising insights, their practical usability remains constrained. On the one hand, the current research landscape predominantly revolves around major sports disciplines such as American football [43] or soccer [16], leveraging the abundance of available data. However, there remains a notable gap in addressing the specific needs of sports education settings that often rely on small player pools. In other words, there is a gap in *educational* data mining research for sports.

Additionally, empirical assessments that explain matchmaking to the players themselves are lacking. This is a critical oversight, as user-centric evaluations are important to build trust and acceptance for decision support systems [4], especially in educational contexts. Our current work aims to address these gaps by presenting an approach for 2v2 matchmaking in small player pools that is extended by an interpretable user interface, and validated on synthetic data as well as in the field.

As with many matchmaking schemes, our proposed algorithm requires as input a knowledge tracing system. Knowledge tracing, under the label of "skill rating", has a long tradition in sports, beginning with counting past victories and losses (*accumulative* rating) and continuing with *adjustive* rating systems like Elo and glicko in chess [11, 12, 35]. Adjustive rating systems aim to model the skill of a player i at time t with one number $\theta_{i,t}$ and the probability $p_{i,j,t}$ of winning against another player j via

$$p_{i,j,t} = \frac{1}{1 + \exp(-\beta \cdot [\theta_{i,t} - \theta_{j,t}])}, \tag{1}$$

where the slope β depends on the specific rating system. After each game, θ_i is updated by adding $K \cdot (y_{i,j,t} - p_{i,j,t})$, where k is a system-specific factor and $y_{i,j,t}$ is the outcome of the game (1 for win, 0 for loss). In the classic FIDE variant of the Elo system, $K = 10$ and $\beta = \log(10)/400$, with ratings starting at $\theta_{i,0} = 1500$ [11]. Our approach is agnostic regarding the specific choice of rating system. In our experiments, we employ a variant of glicko2 [13].

The adjustive rating systems listed above only handle 1v1 games with binary outcomes. By contrast, we consider 2v2 games with a multi-valued outcome, namely the difference in points between the winning and losing teams (margin of victory). Kovalchik [20] reviews several extensions of adjustive rating systems to margin-of-victory outcomes. In this work, we employ the logistic transformation. Further, Williams [41] reviews a number of extensions of adjustive rating systems for teams. We use the composite team approach, i.e., we build a composite skill rating for each team from the ratings of the participating players and then distribute the resulting skill update to the team members.

Once ratings are computed, we can use them for matchmaking. Early work in this domain has focused on matchmaking for tournaments, such as the Swiss system for chess [46]. Most recent work is focused on matchmaking for online video games, where logged-in players shall be matched against other players of similar skill while minimizing the amount of waiting time until a game starts [26]. To our knowledge, no matchmaking algorithm to date has addressed the problem of 2v2 matchmaking in small player pools. That is the challenge we address in this work.

We propose to take an educational data mining view on the matchmaking problem and consider it as an optimization task, using knowledge tracing models as input. Several prior works in educational data mining have focused on optimizing the distribution of students into groups [2, 9, 45]. These works assume that students work collaboratively on a group project, where a high diversity of skills and similarity of interests may be the objective [9, 45]. Empirical investigations assessing the efficacy of grouping strategies in educational settings suggest a nuanced relationship between group composition in terms of individual ability and student performance. On the one hand, there appears to be a tendency for low-ability students to perform better in heterogeneous groups featuring a variety of ability levels [44, 32]. On the other hand, educational settings that group students according to their skill level have shown to improve individual academic success [19], with particular benefits for average- [32] and high–performing [5] students. In line with these empirical insights, Agrawal et al. [2] emphasize the importance of minimizing skill differences while maintaining some variation in ability. This approach allows less skilled team members to learn from their more advanced peers, whereas more skilled team members can learn by teaching.

3. METHOD

Our goal is to distribute a set of n students into teams of two and match opposing teams, such that a) the skill gap between the strongest and weakest player in a match is small, b) skill differences between opposing teams are small, and c) teammates and opponents vary. More precisely, we define a *match* as a vector $\vec{m} = (i, j, r, s)$ of four different player indices in the range $1, \ldots, n$, where i and j indicate the players on the first team and r, s indicate the players on the second team. Further, let $C : \{1, \ldots, n\}^4 \to \mathbb{R}$ be a function that quantifies the *cost* of a match, in the sense that it quantifies how much our three constraints a), b), and c) are violated. Then, our formal optimization problem is given as

$$\begin{aligned} \min_{\vec{m}_1,\ldots,\vec{m}_K \in \{1,\ldots,n\}^4} \quad & \sum_{k=1}^{K} C(\vec{m}_k) && \\ \text{such that} \quad & |\text{set}(\vec{m}_k)| = 4 && \forall k \\ & \text{set}(\vec{m}_k) \cap \text{set}(\vec{m}_l) = \emptyset && \forall k \neq l, \end{aligned} \tag{2}$$

where $\text{set}(i, j, r, s) := \{i, j, r, s\}$ is an operator to translate matches into sets and $K \leq n/4$ is the number of matches we want to be played. In other words, we wish to partition the players onto K matches (potentially with some players

left over), such that the sum over the cost of all matches is minimized and no player participates in multiple matches. We call this the *2v2 matchmaking problem*.

In the following, we first formally define our cost function, then prove that the 2vs2 matchmaking problem is NP-hard, and finally present our solution approach, namely translating the 2v2 matchmaking problem into an integer linear program for which efficient heuristics are available.

3.1 Cost Function

Our cost function has three parts. a) We wish to punish the skill gap between the strongest and weakest player in a match. More precisely, given a match (i,j,r,s), we define $C^a(i,j,r,s) = \max\{\theta_i,\theta_j,\theta_r,\theta_s\} - \min\{\theta_i,\theta_j,\theta_r,\theta_s\}$. b) We wish to punish skill differences between teams. We do so via the difference in team ratings. We define the team rating as two-thirds of the higher rating inside the team plus one-third of the lower rating inside the team, i.e. $\theta_{ij} := \frac{2}{3}\max\{\theta_i,\theta_j\} + \frac{1}{3}\min\{\theta_i,\theta_j\}$. Our hypothesis is that the stronger player in a team affects the team's performance more, hence the higher weight. Additionally, we performed pilot testing with different weightings and found that this weighting predicted game outcomes best. The cost, then, is defined as $C^b(i,j,r,s) = |\theta_{ij} - \theta_{rs}|$. c) We wish to punish if the same people play together. In general, we punish a repetition of teammates with a value of 64, and a repetition of opponents (or crossover from opponent to teammate and vice versa) with a value of 16. For each step into the past, we discount these values with a factor of 0.5 and set values below 1 to zero. For example, if match (i,j,r,s) occurs in round 1 and match (i,x,j,y) in round 2, then, in round 3, match (i,j,u,v) would have cost $C^c(i,j,u,v) = 16 + \frac{1}{2}\cdot 64$. The values and decays can be adjusted depending on how strictly and how long repetitions should be avoided. Generally speaking, the smaller the player pool, the smaller the decay factor should be to permit repetitions earlier again.

Next, we pre-process all costs to magnify values that are above a certain, adjustable threshold. This pre-processing is defined as $\phi(x,\epsilon) = 0.01\cdot x$ if $x < \epsilon$ and $\phi(x,\epsilon) = x - 0.99\cdot\epsilon$, otherwise (so-called leaky rectified linear unit). Finally, we set our overall cost function to

$$\begin{aligned} C(\vec{m}) = \quad & w^a \cdot \phi\big(C^a(\vec{m}),\epsilon^a\big)^2 \\ & + w^b \cdot \phi\big(C^b(\vec{m}),\epsilon^b\big)^2 \\ & + w^c \cdot \phi\big(C^c(\vec{m}),\epsilon^c\big)^2, \end{aligned} \tag{3}$$

meaning a weighted sum of squares after pre-processing. In our experiments, the weights w^a, w^b, and w^c, as well as the thresholds ϵ^a, ϵ^b, and ϵ^c were manually set by the instructor after pilot testing (Table 1). However, our approach is agnostic to the specific parameter settings or the choice of cost function. Future work may consider omitting the pre-processing, setting the weights automatically, or investigate multi-objective optimization to avoid weights altogether.

3.2 Integer Linear Programming Approach

Given the cost function, the 2v2 matchmaking problem (2) is fully specified. If this were a 1v1 matchmaking problem, the problem could now be solved efficiently via the Hungarian algorithm. Unfortunately, though, the 2v2 matchmaking problem is NP-hard, meaning there is no known way to solve it efficiently. We can prove the NP-hardness by reduction of the 4-partition problem onto our problem.

Table 1: The weights w^a, w^b, w^c and thresholds ϵ^a, ϵ^b, ϵ^c for the three cost criteria as used in Eq. (3).

criterion	weight w	threshold ϵ
a) skill gap	1	23
b) balance	0.45	199
c) variety	2.25	0

THEOREM 1. *The 2v2 matchmaking problem is NP-hard.*

PROOF. *Refer to Appendix A.* □

Given that our problem is NP-hard, only heuristics are possible. To obtain such heuristics, we re-write the 2v2 matchmaking problem as an integer linear program (ILP). For ILPs, a host of efficient approximation heuristics exist, especially for binary ILPs [25, 42]. We propose a particular ILP formulation for the 2v2 matchmaking problem that already permits us to perform substantial pre-processing, hence simplifying the solution. In particular, let $\mathcal{M}_1,\ldots,\mathcal{M}_Q$ be all possible subsets of exactly four players from our overall set. Note that Q is n choose 4 or $n\cdot(n-1)\cdot(n-2)\cdot(n-3)/24$. Therefore, the subsets $\mathcal{M}_1,\ldots,\mathcal{M}_Q$ are still efficiently precomputable for moderate choices of n (in our case, we consider only cases up to $n = 40$ players). Further, let c_q be the minimum cost that can be achieved by arranging the players in $\mathcal{M}_q$ into matches. More precisely, if $\mathcal{M}_q = \{i,j,r,s\}$, then $c_q = \min\{C(i,j,r,s), C(i,r,j,s), C(i,s,j,r)\}$. Finally, let $\boldsymbol{A}$ be an $n\cdot Q$ matrix, where $a_{i,q} = 1$ if $i \in \mathcal{M}_q$ and $a_{i,q} = 0$, otherwise. Then, we obtain the following ILP:

$$\begin{aligned} \min_{\vec{x}\in\{0,1\}^Q} \quad & \vec{c}^T \cdot \vec{x} \\ \text{such that} \quad & \boldsymbol{A}\cdot\vec{x} \le \vec{1}, \\ & \vec{1}^T\cdot\vec{x} = K, \end{aligned} \tag{4}$$

where $x_k = 1$ expresses that the qth subset is part of our solution. This ILP is equivalent to the 2v2 matchmaking problem because: 1) The objective function $\vec{c}^T\cdot\vec{x}$ sums the cost of all matches in our solution and is thus equivalent to the objective function in (2). 2) The multiplication of the ith row of $\boldsymbol{A}$ and $\vec{x}$ represents the number of matches in our solution that player i is part of. Hence, the first side constraint ensures that no player participates in more than one match. 3) The second side constraint ensures that exactly K matches are selected.

3.3 User Interface and Explanations

To communicate the matchmaking result to the players, we use a user interface with three parts (refer to Figure 1). The first part displays the matches for the next round (top left), directly guiding players' actions. The second part displays the current ratings of the players (bottom left), thus explaining the matches in terms of rating differences. The final part provides a variant of a counterfactual explanation,

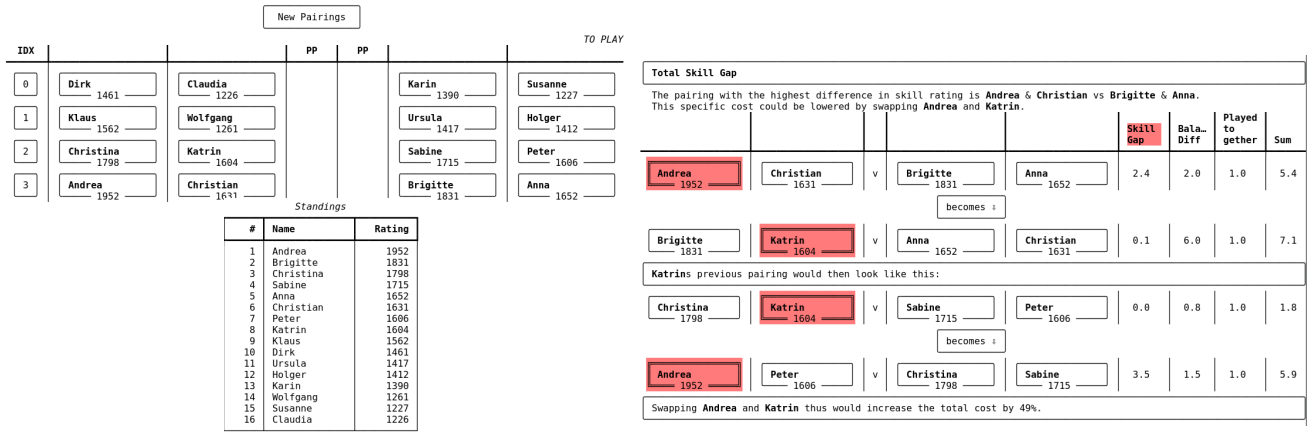

Figure 1: The components of the user interface of our proposed matchmaking system. Note that the displayed names are fictional to preserve anonymity. Top Left: The matches for the next round. Bottom left: The current ratings of the players. Right: A counterfactual explanation, where two players are swapped, resulting in worse costs overall.

which have been shown to provide greater user satisfaction and trust compared to causal explanations [8, 40] (right). In particular, we consider the current worst match in terms of one of the cost criteria and consider the counterfactual case of switching one player to another match to improve the cost. However, we also display how this change necessarily increases the overall cost, either by worsening the cost of another criterion for the original match or by increasing the cost of the match the player was switched to. This design choice enhances usability, as prior work indicates greater usability of counterfactuals when they exhibit minimal changes [21].

4. EXPERIMENTS

We present two experiments. Our first experiment compares our proposed ILP-based solution approach against several baselines on synthetic data. Our second experiment is a field study of the integrated skill tracking, matchmaking, and explanation system in an actual Roundnet course. The source code of our experiments can be found at `https://github.com/sruettgers/automatic_matchmaking`.

4.1 Algorithmic Evaluation

In this experiment, we wish to investigate two questions: A) Does our suggested ILP solution approach yield lower cost function values compared to baseline approaches, and B) how does our proposed approach scale with class size in terms of runtime?

Experimental Setup: For each matchmaking algorithm, we simulated 10 runs of players starting with normally distributed skill values, then being matched by the algorithm, playing against each other, gaining or losing skill as determined by the glicko2 algorithm, and being matched again for four consecutive rounds. This yielded 40 matchmaking rounds overall. For each matchmaking round, we evaluate the cost functions C^a, C^b, and C^c from Section 3.1, as well as the overall cost C in (3). The weights w^a, w^b, w^c and thresholds ϵ^a, ϵ^b, ϵ^c chosen in the experiments are listed in Table 1. All experiments were performed on a consumer-grade desktop PC with an AMD Ryzen 5 1600 CPU (2017) and 32 GB RAM.

Algorithms: We considered the following matchmaking algorithms: **Random** permutes the list of players randomly and then assigns the first four players to the first match, the next four to the next match, and so on. This approach is likely to perform well in terms of C^c, because players vary a lot, but badly in terms of C^a and C^b. **Fixed order** sorts the players according to their skill rating and then assigns the first four players to the first match, the next four to the next match, and so on. This approach is likely to perform well in terms of C^a and C^b, because skill differences are likely to be small, but badly in terms of C^c, because the same players are likely to be matched each time. **DoubleSymmetric** optimizes the assignment of teammates and the assignment of teams independently of each other. Each is a classic assignment problem, which can be efficiently solved via the Hungarian algorithm [22]. The problem is that the team assignment may limit the abilities of the algorithm to match balanced opponents, such that the overall difference in maximum and minimum skill C^a, as well as the skill difference between teams C^b may be negatively affected. **MiniZinc** is a formal language for constraint programming [30]. We expressed the 2v2 matchmaking problem (2) in MiniZinc and then applied the `gecode` solver. **ILP_SciPy** refers to our proposed ILP formulation (4), solved via the solver in the scipy package [38]. **ILP_MiniZinc** refers to our proposed ILP formulation (4), solved via the `COIN-BC` solver in MiniZinc.

Results: Figure 2 displays the overall costs we obtained with each approach as box plots. We observe that the cost function values obtained via ILP approaches were significantly lower compared to the other approaches ($p < 0.05$ in a Wilcoxon signed rank test), but that both SciPy and MiniZinc achieved roughly the same results on the ILP formulation (4). Table 2 lists the results for the single cost criteria. As expected, random achieves the best results in terms of the variety criterion C^c, but the worst results on the

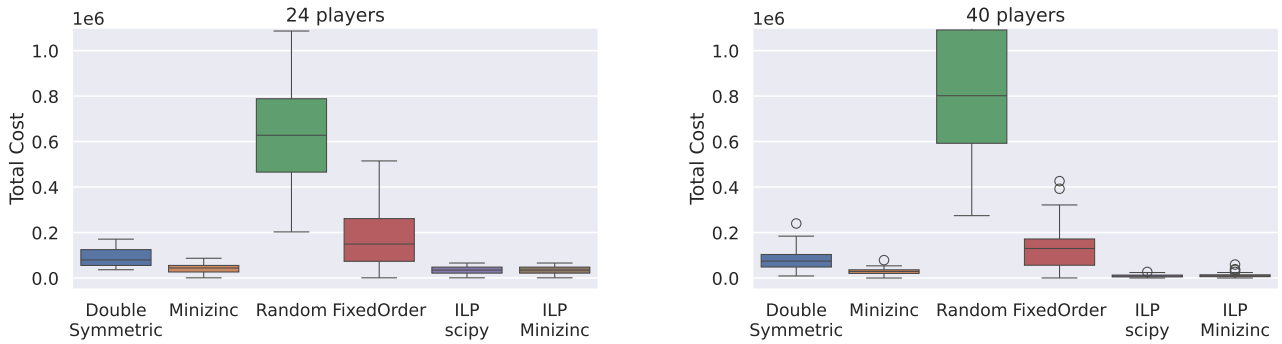

Figure 2: The overall cost (3) for the matches generated via each matchmaking algorithm for 24 players (left) and 40 players (right).

Table 2: Mean pre-processed and squared cost for each cost criterion $\phi(C^a, \epsilon^a)^2$, $\phi(C^b, \epsilon^b)^2$, and $\phi(C^b, \epsilon^b)^2$ as well as the overall cost C from Eq. (3) for each matchmaking algorithm for the 40 simulated 24 player rounds. The weights w^a, w^b, w^c and thresholds ϵ^a, ϵ^b, ϵ^c are listed in Table 1.

algorithm	$\phi(C^a,\epsilon^a)^2$	$\phi(C^b,\epsilon^b)^2$	$\phi(C^c,\epsilon^c)^2$	C
Random	212744	390424	26730	629899
FixedOrder	**3433**	8293	170560	182286
DoubleSymm.	13941	55305	19848	89094
Minizinc	12545	6581	22362	41487
ILP_scipy	16433	**4677**	**12216**	**33326**
ILP_Minizinc	16433	**4677**	**12216**	**33326**

skill difference criteria C^a and C^b. Conversely, FixedOrder achieves the best values in terms of C^a but performs particularly poorly in terms of variety C^c. DoubleSymmetric does not perform particularly well on any measure, whereas Minizinc performs reasonably close to optimal but is outperformed by the ILP results in C^b and C^c.

Figure 3 displays box plots for the runtimes of all algorithms. We observe that, as expected, the random and fixed order baselines were the fastest, with DoubleSymmetric close behind. For 24 players, ILP_SciPy is the third-fastest but becomes by far the slowest solver for 40 players. This is due to an approximately exponential runtime scaling. Overall, the ILP_Minizinc solver appears to be the most reliable choice to achieve the best results while maintaining a fast runtime.

4.2 Field Study

In our field study, we compared the proposed matchmaking approach with a random baseline and evaluated user satisfaction with the match quality, as well as the perceived goodness, efficacy, and trustworthiness of the provided explanation utility.

Roundnet: The study was performed in a Roundnet course at Bielefeld University, Germany. Roundnet is a 2v2 sport played with a round net in the center. One team starts passing the ball with up to three alternating touches before the ball has to be played onto the net and the other team's turn begins. If a team fails to play the ball onto the net within three touches, the opposing team receives a point. The round ends once a team reaches 15 points.

Study Setup: Participants ($N = 38$) were recruited during four consecutive sessions of a weekly Roundnet course in a sports hall at Bielefeld University. The course was open to all university students and employees. While demographic data was not recorded to avoid the risk of making survey responses re-identifiable, the course population were mostly students in an age range 19-25. Participants were free to participate in neither, one ($n = 20$), two ($n = 3$), three ($n = 6$), or all four ($n = 9$) sessions. Two sessions used the proposed matchmaking system, two the random baseline (unknown to the participants).

Each session lasted 90 minutes. At the start, the instructor explained the study and handed out information sheets. Next, automatic matchmaking (using the proposed approach or the random baseline) was performed and the generated matches with explanations were displayed on a 44" computer screen (refer to Section 3.3). After matches were completed, the instructor entered the results into a laptop and skill parameters were updated using glicko2 [13] with logistic transformation to account for the margin of victory [20] and composite team extension to distribute rating adjustments to single team members [41]. The matchmaking, playing, and rating steps were repeated for three additional rounds. After the four rounds of Roundnet, participants filled out a paper survey and were compensated with 20 EUR. The study was approved beforehand by the ethics board of Bielefeld University.

Evaluation Measures: The paper survey included the Explanation Goodness Checklist and the Trust Scale, both proposed by Hoffman et al. [14], and an adapted version of the System Causability Scale by Holzinger et al. [15]. Specifically, we excluded the item "I could change the level of detail on demand" in our version because this item did not apply. Further, to evaluate participants' subjective impressions of the matching quality, we included two items per cost criterion from Section 3.1. All items (and full survey results) are listed in Appendix B. The significance of differences between groups was evaluated using a non-parametric Mann-Whitney U test.

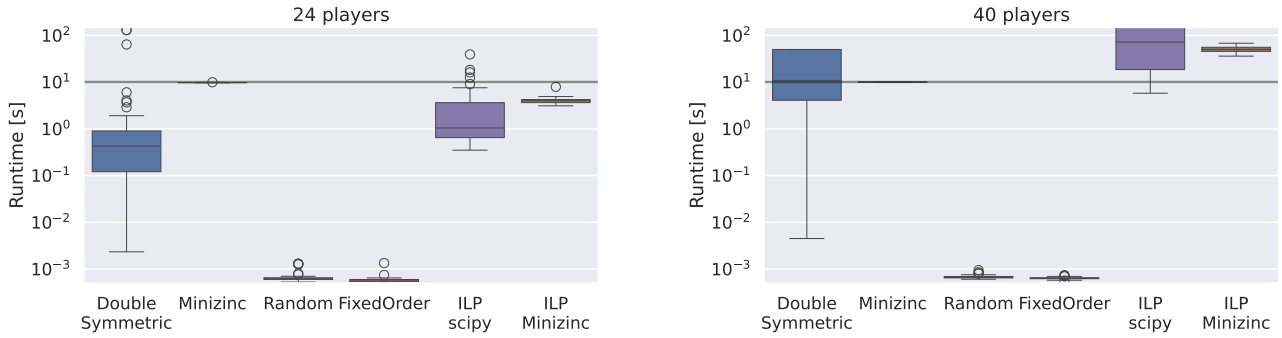

Figure 3: The runtime of all matchmaking algorithms for 24 players (left) and 40 players (right) for a single round.

Results: The full results across all survey items are listed in Appendix B. We provide a summary of the most important results in the following.

On the explanation goodness checklist, participants responded favorably on most items in both conditions.In terms of statistically significant differences between groups, participants in the optimized condition agreed significantly more that the explanations were actionable (89% in the optimized condition vs. 71% in the random condition) and helped them to understand how the tool works (91% in the optimized condition vs. 68% in the random condition).

On the trust scale, participants overwhelmingly agreed or strongly agreed that the tool could perform the task better than a human novice (88% in the optimized condition, and 77% in the random condition) and was efficient (86% in the optimized condition, and 94% in the random condition). 57% of participants in the optimized condition, and 42% of participants in the random condition indicated agreement or strong agreement toward the items assessing the reliability of the tool. This strong indication of agreement between scales instills confidence in the consistency and reliability of the overall results. Results on the item asking for predictability of the tool were less positive, again to a comparable extent between both groups (75% in the optimized group, 73% in the random group): in both cases, a majority in both groups gave either "neutral", "disagree", or "strongly disagree" responses concerning the notion that the tool was very predictable. The statistical analysis of differences between groups revealed a more positive evaluation of the optimized matchings compared to random. The participants' responses indicated that they liked using the system more in the optimized condition, were less wary of it, felt more safe, and more confident in the tool compared to the random condition.

On the System Causability Scale, again, participants overwhelmingly judged both conditions favorably, in terms of explanation presentation, sort of information provided, and helpfulness of explanations to understand causality. The majority of participants in both groups also agreed with the notion that most people would learn to understand the explanations quickly. The only item eliciting a more balanced distribution (45% neutral and 15% disagree or strongly disagree for the optimized condition; 19% neutral and 27% disagree or strongly disagree for the random condition) concerned the extent to which the presented data included all relevant known causal factors with sufficient precision and granularity, which is, arguably, the most demanding. In terms of significant effects, partcipants rated the optimized condition as significantly more consistent (84% agreement or strong agreement) compared to the random condition (59% agreement or strong agreement). Similarly, participants found the explanations in the optimized condition significantly more consistent and usable with their knowledge base (89% agreement or strong agreement in the optimized condition vs. 72% agreement or strong agreement in the random condition).

In terms of subjective quality of the matchmaking, participants agreed that all players as well as opposing players had similar skill in their matches for the optimized condition (45% agreement or more) but not in the random condition (58% disagreement or more). The difference between conditions was significant. In terms of player variability, the difference between conditions was less pronounced.

5. DISCUSSION AND CONCLUSION

In this paper, we proposed an educational data mining solution to the matchmaking problem in 2v2 sports. In particular, we propose to trace players' knowledge via glicko2 and then perform the matchmaking based on the estimated skill values for all players. We showed that, in contrast to 1v1 sports, the 2v2 matchmaking problem is NP-hard, but that efficient heuristics are possible by rephrasing the problem as an integer linear program (ILP). In simulation studies, we showed that this ILP formulation yielded matches with significantly lower costs while remaining acceptable in terms of runtime, at least for up to 40 players.

In a field study with $N = 38$ participants, we evaluated an integrated system of knowledge tracing, matchmaking, and counterfactual explanations of the matchmaking, in comparison to a random baseline with the same user interface. It is noteworthy that participants generally perceived explanations positively, even in the random condition. This not only indicates that the user interface and explanation design were indeed suitable (see Section 3.3). It also underscores the effectiveness of the selected explanation format,

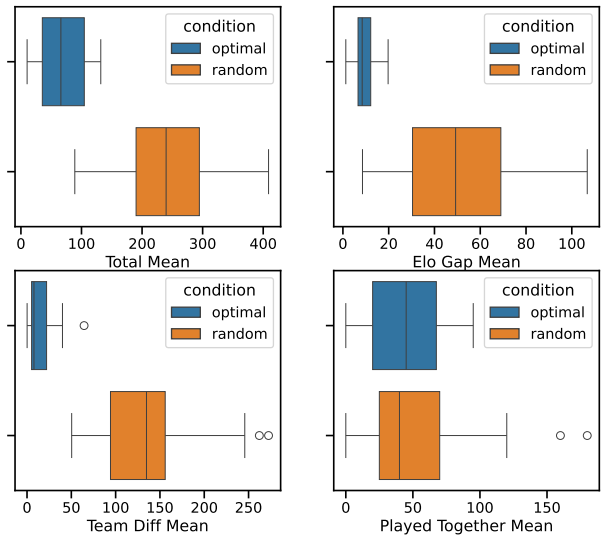

Figure 4: Actual cost function values for the optimized and random conditions in the field study. Top left: overall cost. Top right: Rating difference between strongest and weakest player (a). Bottom left: Rating difference between opposing teams (b). Bottom right: Player variance (c).

i.e., counterfactual explanations, aligning with the widely accepted notion that this approach offers users valuable insights into systems by closely resembling the way humans themselves reason [8]. Even in the random condition, participants found the explanations to be complete, sufficiently detailed, and satisfying (Fig. 5), showcasing the intuitive access provided by them. This aspect, while seemingly beneficial, warrants caution: counterfactuals may, paradoxically, be misleading due to their intuitive appeal [24]. Indeed, participants in both groups gave comparable judgments on a subset of items in the Trust scale (refer to Figure 6), indicating that they perceived the tool to have superior performance than a novice user and to work efficiently, independent of the underlying matchmaking mechanism. While the absence of a group effect on these items defies conclusive interpretation, it may strengthen the hypothesis that explainability fosters misplaced trust in a less-than-ideal system.

However, we also find multiple significant differences between conditions, suggesting that users *can* recognize differences between the underlying systems. In particular, users were significantly less wary of the proposed matchmaking system, liked to use it more for decision making, felt more safe relying on the tool, were more confident that it worked well, found the explanations more consistent, and felt more able to use the explanations with their knowledge base. Accordingly, there is some evidence that users are indeed able to recognize whether the system underlying the explanations warrants their trust. We found that participants recognized that the proposed matchmaking system reduced skill differences inside matches and skill differences between opposing teams significantly compared to the random baseline. We note, however, that many ratings were also favorable in the control condition, indicating that the mere fact of having automatic matchmaking with intuitive and easily accessible explanations was already deemed helpful.

Finally, participants recognized differences in matchmaking quality between optimized and random condition: The proposed system achieved considerably smaller skill gaps within matches and between opposing teams (refer to Figure 4), which was recognized by the participants (refer to Figure 8). However, both conditions achieved variability between teammates and opponents, which the subjective ratings by participants reflect (i.e. there was no significant difference in perceived player variability between conditions).

Limitations: We believe that our experiments show encouraging results and indicate that the proposed matchmaking approach is applicable for actual matchmaking in 2v2 sports. Nonetheless, some limitations remain. First, some of our cost-function design choices may be specific to our setting and may need to be changed in other contexts. Second, we only explored a glicko2 system for knowledge tracing. Other knowledge tracing approaches may be more suitable (e.g. yield more accurate predictions). Third, we focused on counterfactual explanations for the matchmaking, while other types of explanations (e.g. of the skill ratings) may also be helpful. Fourth, our study was too short to meaningfully investigate whether matchmaking had a positive effect on skill acquisition. Future work should perform longitudinal studies to compare the learning effect of automatic matchmaking.

Impact and Ethics: We hope that our proposed matchmaking algorithm provides a varied set of teammates and opponents of comparable skill for all players, thus enhancing their skill acquisition in 2v2 sports as well as their enjoyment of the game. Nonetheless, practitioners should also be aware of potential drawbacks. By making skill ratings very transparent during the entire training process, players may get more competitive and more focused on improving their ranking, rather than their actual skill or being good teammates and opponents. More broadly, the extrinsic motivation of improving the rating may replace intrinsic motivation and thus hurt learning as well as enjoyment of the game. In such cases, randomized matchmaking may be preferable.

Still, in our present investigation, participants consistently perceived our method as a valuable, actionable, and useful tool for matchmaking in the context of a sports course. The validation of our tool in a real-life setting greatly enhances its value, affirming its effectiveness and dependability in practical scenarios. Thus, our work extends the research landscape by presenting a decision support system for skill rating and matchmaking in the realm of sports education, a context that presents unique challenges such as small player pools and the absence of large pre-existing databases.

6. ACKNOWLEDGMENTS

U.K. was supported by the research training group “Dataninja” (Trustworthy AI for Seamless Problem Solving: Next Generation Intelligence Joins Robust Data Analysis) funded by the German federal state of North Rhine-Westphalia.

7. REFERENCES

[1] G. Abdelrahman, Q. Wang, and B. Nunes. Knowledge tracing: A survey. *ACM Computing Surveys*, 55, 2023.

[2] R. Agrawal, B. Golshan, and E. Terzi. Grouping students in educational settings. In S. Macskassy and C. Perlich, editors, *Proceedings of the 20th ACM SIGKDD*, page 1017–1026, 2014.

[3] L. Babel, H. Kellerer, and V. Kotov. The k-partitioning problem. *Mathematical Methods of Operations Research*, 47:59–82, 1998.

[4] T. A. Bach, A. Khan, H. Hallock, G. Beltrão, and S. Sousa. A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human–Computer Interaction*, pages 1–16, 2022.

[5] J. Baer. Grouping and achievement in cooperative learning. *College teaching*, 51(4):169–175, 2003.

[6] J. D. Bartlett, F. O'Connor, N. Pitchford, L. Torres-Ronda, and S. J. Robertson. Relationships between internal and external training load in team-sport athletes: evidence for an individualized approach. *International journal of sports physiology and performance*, 12(2):230–234, 2017.

[7] R. P. Bunker and F. Thabtah. A machine learning framework for sport result prediction. *Applied computing and informatics*, 15(1):27–33, 2019.

[8] R. M. Byrne. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In S. Kraus, editor, *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6276–6282, 2019.

[9] M.-I. Dascalu, C.-N. Bodea, M. Lytras, P. O. de Pablos, and A. Burlacu. Improving e-learning communities through optimal composition of multidisciplinary learning groups. *Computers in Human Behavior*, 30:362–371, 2014.

[10] DOSB. Bestandserhebung 2023. Technical report, German Olympic Sports Confederation, 2023. [German].

[11] A. E. Elo and S. Sloan. *The rating of chessplayers: Past and present*. American Chess Foundation, New York, NY, USA, 1978.

[12] M. E. Glickman. The glicko system. Technical report, Boston University, 1995.

[13] M. E. Glickman, J. Hennessy, and A. Bent. A comparison of rating systems for competitive women's beach volleyball. *Statistica Applicata - Italian Journal of Applied Statistics*, 30(2):233–254, 2018.

[14] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable AI: challenges and prospects. *CoRR*, abs/1812.04608, 2018.

[15] A. Holzinger, A. Carrington, and H. Müller. Measuring the quality of explanations: the system causability scale (scs) comparing human and machine explanations. *KI-Künstliche Intelligenz*, 34(2):193–198, 2020.

[16] A. Jarvandi, S. Sarkani, and T. Mazzuchi. Modeling team compatibility factors using a semi-markov decision process: A data-driven approach to player selection in soccer. *Journal of Quantitative Analysis in Sports*, 9(4):347–366, 2013.

[17] A. Jaspers, T. O. De Beéck, M. S. Brink, W. G. Frencken, F. Staes, J. J. Davis, and W. F. Helsen. Relationships between the external and internal training load in professional soccer: what can we learn from machine learning? *International journal of sports physiology and performance*, 13(5):625–630, 2018.

[18] S. Jauhiainen, S. Äyrämö, H. Forsman, and J. Kauppi. Talent identification in soccer using a one-class support vector machine. *International Journal of Computer Science in Sport*, 18(3), 2019.

[19] B. Khazaeenezhad, H. Barati, and M. Jafarzade. Ability grouping as a way towards more academic success in teaching EFL–a case of iranian undergraduates. *English Language Teaching*, 5(7):81–89, 2012.

[20] S. Kovalchik. Extension of the Elo rating system to margin of victory. *International Journal of Forecasting*, 36(4):1329–1341, Oct. 2020.

[21] U. Kuhl, A. Artelt, and B. Hammer. Keep your friends close and your counterfactuals closer: Improved learning from closest rather than plausible counterfactual explanations in an abstract setting. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2125–2137, 2022.

[22] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[23] M. Lai, R. Meo, R. Schifanella, and E. Sulis. The role of the network of matches on predicting success in table tennis. *Journal of sports sciences*, 36(23):2691–2698, 2018.

[24] H. Lakkaraju and O. Bastani. "How do i fool you?" manipulating user trust via misleading black box explanations. In A. Markham, J. Powles, T. Waslh, and A. Washington, editors, *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85, 2020.

[25] J.-H. Lange and P. Swoboda. Efficient message passing for 0–1 ILPs with binary decision diagrams. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, pages 6000–6010, 2021.

[26] Y. Liu, Y. Qi, J. Zhang, C. Kou, and Q. Chen. MMBench: The Match Making Benchmark. In T.-S. Chua and H. Lauw, editors, *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1128–1131, 2023.

[27] F. Louzada, A. C. Maiorano, and A. Ara. isports: A web-oriented expert system for talent identification in soccer. *Expert Systems with Applications*, 44:400–412, 2016.

[28] G. Morciano, A. Zingoni, A. Morachioli, and G. Calabrò. Machine learning prediction of the expected performance of football player during training. In P. Arpaia and L. T. De Paolis, editors, *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, pages 574–578. IEEE, 2022.

[29] A. Naglah, F. Khalifa, A. Mahmoud, M. Ghazal, P. Jones, T. Murray, A. S. Elmaghraby, and

A. El-Baz. Athlete-customized injury prediction using training load statistical records and machine learning. In E. Abdel-Raheem and M. Tebaldi, editors, *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 459–464. IEEE, 2018.

[30] N. Nethercote, P. J. Stuckey, R. Becket, S. Brand, G. J. Duck, and G. Tack. MiniZinc: Towards a Standard CP Modelling Language. In C. Bessière, editor, *Proceedings of the 2007 Conference on Principles and Practice of Constraint Programming*, pages 529–543, 2007.

[31] J. D. Ruddy, A. J. Shield, N. Maniar, M. D. Williams, S. J. Duhig, R. G. Timmins, J. Hickey, M. N. Bourne, and D. A. Opar. Predictive modeling of hamstring strain injuries in elite australian footballers. *Medicine & Science in Sports & Exercise*, 50(5):906–914, 2018.

[32] M. Saleh, A. W. Lazonder, and T. De Jong. Effects of within-class ability grouping on social interaction, achievement, and motivation. *Instructional Science*, 33:105–119, 2005.

[33] D. Siedentop, P. Hastie, and H. Van der Mars. *Complete guide to sport education*. Human Kinetics, Champaign, IL, USA, 3 edition, 2019.

[34] J. Song and J. Zhong. Fuzzy theory in the prediction of athletes' competitive state based on information security. *Mobile Information Systems*, 2022, 2022.

[35] R. Stefani. The Methodology of Officially Recognized International Sports Rating Systems. *Journal of Quantitative Analysis in Sports*, 7(4), Oct. 2011.

[36] Z. Taha, R. M. Musa, A. P. A. Majeed, M. M. Alim, and M. R. Abdullah. The identification of high potential archers based on fitness and motor ability variables: A support vector machine approach. *Human movement science*, 57:184–193, 2018.

[37] H. Van Eetvelde, L. D. Mendonça, C. Ley, R. Seil, and T. Tischer. Machine learning methods in sport injury prediction and prevention: a systematic review. *Journal of experimental orthopaedics*, 8:1–15, 2021.

[38] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

[39] X. Wang. Research on the evaluation of sports training effect based on artificial intelligence technology. In K. Subramaniam and A. P. Muthuramalingam, editors, *Second International Conference on Algorithms, Microchips, and Network Applications (AMNA 2023)*, volume 12635, pages 249–255. SPIE, 2023.

[40] G. Warren, R. M. J. Byrne, and M. T. Keane. Categorical and continuous features in counterfactual explanations of AI systems. In F. Chen and M. Billinghurst, editors, *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, 2023.

[41] G. J. Williams. Abstracting Glicko-2 for Team Games. Master's thesis, University of Cincinnati, 2013.

[42] L. A. Wolsey. *Integer programming*. John Wiley & Sons, Hoboken, NJ, USA, 2 edition, 2021.

[43] W. Young and G. R. Weckman. A team-compatibility decision support system for the national football league. *International Journal of Computer Science in Sport*, 19(1):60–101, 2020.

[44] M. Zamani. Cooperative learning: Homogeneous and heterogeneous grouping of Iranian EFL learners in a writing context. *Cogent Education*, 3(1):1149959, 2016.

[45] Y. Zheng, C. Li, S. Liu, and W. Lu. An improved genetic approach for composing optimal collaborative learning groups. *Knowledge-Based Systems*, 139:214–225, 2018.

[46] S. Ólafsson. Weighted matching in chess tournaments. *The Journal of the Operational Research Society*, 41(1):17–24, 1990.

APPENDIX

A. NP-HARDNESS PROOF

THEOREM 2. *The 2v2 matchmaking problem is NP-hard.*

PROOF. *Refer to Appendix A. Consider the 4-partition problem: Given n real numbers $w_1, \ldots, w_n$ where n is divisible by four; does there exist a partition of these numbers into $K = n/4$ subsets $S_1, \ldots, S_K$, such that the sum of the numbers in each set is equal? This problem is NP-hard [3]. We can convert such a 4-partition problem into a 2v2 matchmaking problem with n players, $K = n/4$, and cost function $C(i, j, r, s) = \left(w_i + w_j + w_r + w_s\right)^2$. Further, let $W := \sum_{i=1}^{n} w_i$. Our claim is now: the 4-partition problem is solvable if and only if our 2v2 matchmaking problem is solved with the objective function value $K \cdot (W/K)^2$.*

For the forward direction, let $S_1, \ldots, S_K$ be a solution to the 4-partition problem. Then, the numbers in each set must sum to the same number, meaning the sum is W/K in each case. Further, let $S_k = \{w_i, w_j, w_r, w_s\}$ and construct a corresponding match $\vec{m}_k = (i, j, r, s)$. Accordingly, $\vec{m}_1, \ldots, \vec{m}_K$ is a solution to the 2v2 matchmaking problem with objective function value $K \cdot (W/K)^2$.

For the backward direction, consider the following continuous relaxation of the 2v2 matchmaking problem:

$$\min_{x_1, \ldots, x_K \in \mathbb{R}} \quad \sum_{k=1}^{K} x_k^2 \quad \textit{such that} \quad \sum_{k=1}^{K} x_k = W.$$

In other words, x_k represents the sum of the numbers in the kth match. The side constraint ensures that the sum of all numbers remains W and the objective function corresponds to the objective function of the 2v2 matchmaking problem.

This relaxation is a convex optimization problem with the unique global optimum $x_1 = \ldots = x_K = W/K$. Therefore, it is impossible to achieve a lower objective function value than $K \cdot (W/K)^2$ in our 2v2 matchmaking problem and the only way to achieve this value is to have matches where all contained numbers add up to W/K. Now, let $\vec{m}_1, \ldots, \vec{m}_K$ be such a solution. We can translate each match $\vec{m}_k = (i, j, r, s)$ into a set $S_k = \{w_i, w_j, w_r, w_s\}$ and the resulting $S_1, \ldots, S_K$ are by construction a solution to the 4-partition problem.

In other words: If we could solve the 2v2 matchmaking problem efficiently, we could also solve the 4-partition problem efficiently. And because the 4-partition problem is NP-hard, so is the 2v2 matchmaking problem. □

B. FULL SURVEY RESULTS

Figures 5, 6, 7, and 8 show the full results of the survey in the field study. Asterisks denote statistical significance of group differences with $p < .05$.

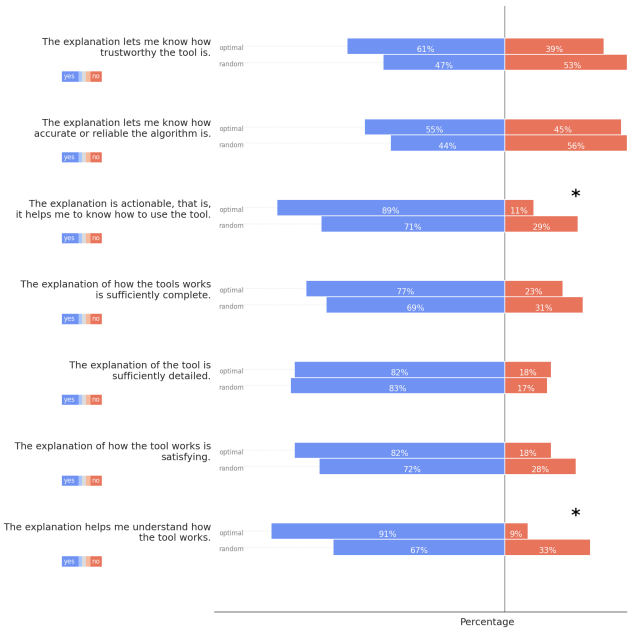

Figure 5: Results of the paper survey for the Explanation Goodness Checklist [14]. Blue bars show the proportion of participants per group responding "Yes", red bars illustrate the proportion of participants responding "No" to a given item.

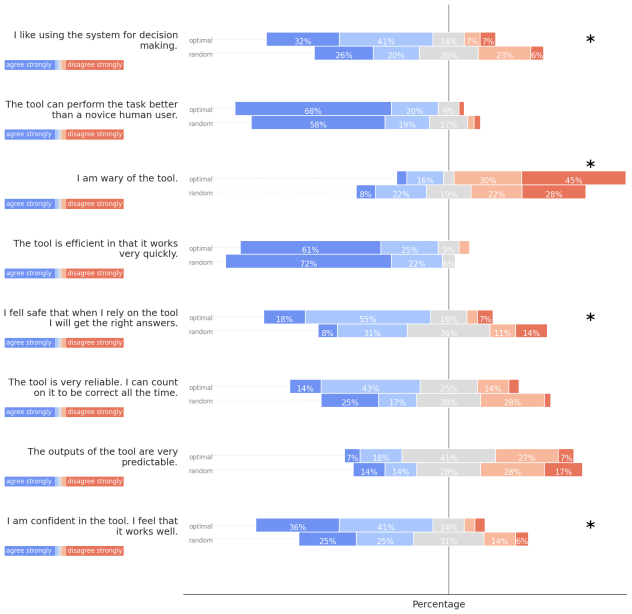

Figure 6: Results of the paper survey for the Trust Scale [14], evaluated on a 5-point Likert scale ranging from "strongly agree" to "strongly disagree". Bars illustrate the proportion of participants responding with either positive (blue shades), neutral (gray), or negative valence (red shades).

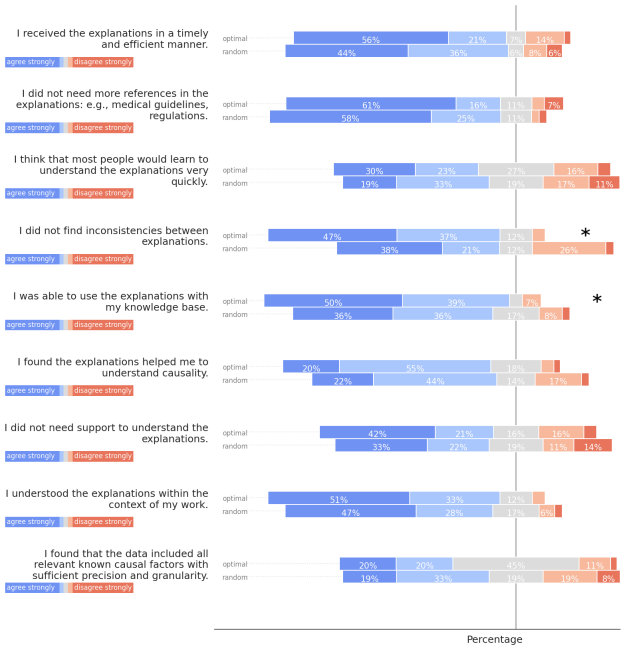

Figure 7: Results of the paper survey for our adapted version of the System Causability Scale[15], evaluated on a 5-point Likert scale ranging from "strongly agree" to "strongly disagree". Bars illustrate the proportion of participants responding with either positive (blue shades), neutral (gray), or negative valence (red shades).

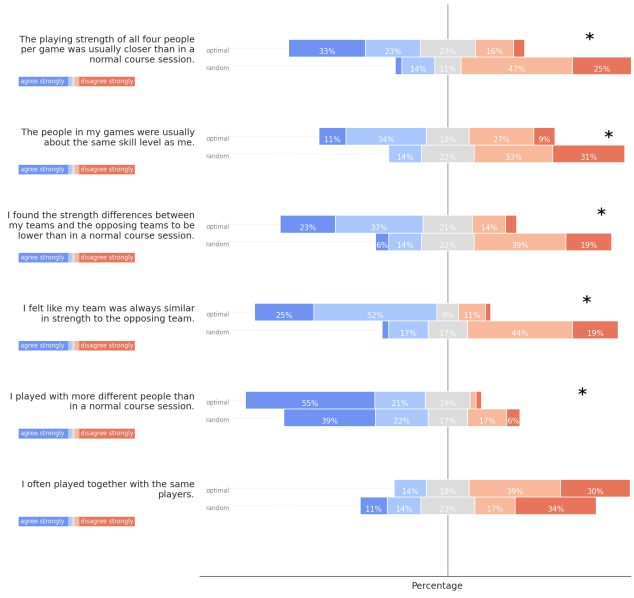

Figure 8: Results of the paper survey for the match quality evaluation, assessed on a 5-point Likert scale ranging from "strongly agree" to "strongly disagree". Bars illustrate the proportion of participants responding with either positive (blue shades), neutral (gray), or negative valence (red shades).