applied sciences

MDPI

Article

JudgED: Comparison between Kickboxing Referee Performance at a Novel Serious Game for Judging Improvement and at World Championships

Dominik Hoelbling [1,2,*,†], Andre Salmhofer [1,†], Cebrail Gencoglu [3,†], René Baranyi [1,4], Karl Pinter [1,2], Serhat Özbay [3], Süleyman Ulupinar [3], Abdullah Bora Ozkara [5] and Thomas Grechenig [1]

[1] Research Group for Industrial Software (INSO), TU Wien, 1040 Vienna, Austria; thomas.grechenig@inso.tuwien.ac.at (T.G.)
[2] Research Industrial Systems Engineering (RISE), Concorde Business Park F, 2320 Schwechat, Austria
[3] Department of Physical Education and Sport, Faculty of Sport Sciences, Erzurum Technical University, Erzurum 25070, Turkey
[4] Human & Digital, 1040 Vienna, Austria
[5] Department of Physical Education, Karadeniz Technical University, Trabzon 61080, Turkey
[*] Correspondence: dominik.hoelbling@inso.tuwien.ac.at
[†] These authors contributed equally to this work.

Abstract: The particular responsibility of referees in combat sports lies in their decision-making to enforce the rules of the sport, which requires considerable experience and a multitude of skills, including perception, categorization, memory processing, and information integration. As a cost-effective alternative to in-tournament training, this research aims to evaluate the novel video-based serious game called "JudgED" to train martial arts referees' decision-making processes through immediate feedback. The effectiveness of the JudgED game was assessed by (a) measuring decision accuracy and specific reaction time, (b) calculating a theoretical probability of correct scoring, and (c) comparing these results with real competition judging agreement data. A field study was conducted to analyze the performance of 16 kickboxing referees. The study involved two video-based tests in the serious game. The performance data for JudgED were obtained via a procedure that compares the players' inputs in the serious game with expert-defined decisions. The results were compared to real-competition data gathered through qualitative analysis of kickboxing fights (n = 400 fights/1200 bouts) at the WAKO World Championships 2021. The findings showed an average decision accuracy of 43.011% and an average reaction time of 1.022 s. For further comparison, binominal distribution for the probability of correct final decisions (between 15.3% and 67.2%) in JudgED and Fleiss' Kappa interrater reliability for JudgED (Ring: $\kappa = 0.371$; Tatami: $\kappa = 0.398$; $p < 0.001$) and tournament decisions (by bout: $\kappa = 0.114$; by fight $\kappa = 0.063$; by outcome $\kappa = 0.166$; $p < 0.001$) were calculated. The results suggest that more training is required to improve referee decision accuracy, and JudgED bears the potential to work as a suitable supporting system.

Keywords: kickboxing; decision-making training; serious game; digital-game-based learning; referees; judges; martial arts; match analysis

check for updates

Citation: Hoelbling, D.; Salmhofer, A.; Gencoglu, C.; Baranyi, R.; Pinter, K.; Özbay, S.; Ulupinar, S.; Ozkara, A.B.; Grechenig, T. JudgED: Comparison between Kickboxing Referee Performance at a Novel Serious Game for Judging Improvement and at World Championships. *Appl. Sci.* **2023**, *13*, 9549. https://doi.org/10.3390/app13179549

Academic Editor: Mark King

Received: 14 July 2023
Revised: 18 August 2023
Accepted: 18 August 2023
Published: 23 August 2023

1. Introduction

Referees must possess diverse skills, including perception, physical fitness, and athlete interaction, but their primary responsibility lies in decision-making [1,2]. In martial arts, where athletes execute rapid sequences of techniques within a short time frame [3,4], referees must rely on their memory to make appropriate decisions. This involves combining their perception of the athletes' movements with their previous experiences and the sport's rules [5,6]. Examining the decision-making task reveals a complex social–cognitive process influenced by various specific external factors inherent to the sport being officiated [7].

To effectively handle this complexity, referees must integrate their declarative knowledge of the sports' rules with procedural knowledge gained through practical experience [8]. Insufficient training of referees can lead to decision errors that have the potential to impact the outcomes of competitions or tournaments [1].

Engaging in martial arts training has been recognized as a sport-based intervention that can contribute to increased well-being [9]. When considering the influence of games in this context, it is worth noting that a study by Jhon [10] has already identified that 3D games have a minor impact on martial arts practice.

1.1. Kickboxing

The sport of kickboxing, according to the rules and regulations of the World Association of Kickboxing Organisations (WAKO) [11], comprises a total of six fighting disciplines, namely the tatami disciplines Point fighting (PF), Light Contact (LC) and Kick Light (KL), as well as the ring disciplines Full Contact (FC), Low Kick (LK) and K1. While in ring disciplines, every regular technique is only scored with 1 point if the judge detects it as an "effective hit", every regular technique leads to 1 to 3 points in the tatami disciplines. For example, all hand techniques and kicks to the body (or leg, if allowed) and leg sweeps are scored with 1 point, while kicks to the head are awarded with 2 points, and jumping kick executions lead to an additional point. Furthermore, it is essential to note that judge decisions include penalties for either exit (stepping out of the fighting area) in tatami disciplines or prohibited techniques or target regions in all disciplines. While most current research on kickboxing competition focuses on time motion analysis [12–14], no scientific work was found on judging performance, which unfortunately generates the necessity of producing additional results in order to compare the game with real-life data.

1.2. Decision-Making Process

Making decisions involves a series of steps within social information processing, which include perception, categorization, memory processing, and information integration [15,16]. The process also occurs in various ways. The literature commonly differentiates between rational and intuitive decisions [17,18]. According to dual-process theories, two types of thought processes exist. One method is swift and instinctive. The other approach involves slower and analytical evaluation [19]. Recent research increasingly discusses automatic decision-making using algorithms [20]. Rational and intuitive decision-making processes are not mutually exclusive; instead, they compliment each other. Combining rational and intuitive processes often offers a solid foundation for decision-making. Bias, however, can influence any decision-making process [21–23] and may also be more prevalent in untrained referees. Furthermore, each step in the decision-making process plays a crucial role in ensuring sound decision-making, although the emphasis on each step may vary depending on the specific characteristics of the situation being judgED [24]. In the specific context of assessing foul/no-foul scenarios in soccer, Schweizer et al. emphasize the importance of the categorization step [25]. They draw upon Brunswik's Lens model [26] to argue that multiple cues influence categorizations, and only relevant cues contribute to the accuracy of the decision. To effectively handle cue integration and make decisions under time constraints, intuitive processing is employed instead of deliberate processing.

1.3. Decision-Making Training Challenges

The investigation of referees' decisions has increased due to their impact on competition outcomes and their associated economic consequences [7,27]. While various approaches have been proposed in the literature to develop referees' decision-making skills, direct participation in sports competitions is recognized as an ideal method for acquiring these skills [28]. However, relying solely on on-field experience may need to provide more training intensity to reach an expert level in decision-making, as suggested by skill development frameworks like the 10,000-hour rule of deliberate practice [29,30].

To address the limited training time resulting from the scarcity of competitive events, video-based training programs have emerged as a potential solution [7]. These programs facilitate the accumulation of practical training intensity that would be difficult to accomplish solely through the judging of real competitions [30]. Over the past 17 years, there has been a growing trend in research towards developing and evaluating well-grounded video-based decision-making training programs [7]. Recent studies have confirmed the effectiveness of video-based training platforms for referees in various sports [25,30–32]. Although there is currently no training platform available specifically for improving the decision-making skills of martial arts referees, the positive effects observed in video-based training approaches could potentially be transferable to martial arts refereeing.

1.4. Serious Games

In order to develope and improve skills with modern technology [33,34], recent research suggests the utilization of serious games for increased motivation and resilience during training [35–37]. Although there are various definitions of the term "serious game", Michael et al. [35] characterize it as games that prioritize education over entertainment. There is a wide range of domains where serious games might be used, like within the rehabilitation process of stroke using the sensors of a smartphone [36], or using smartphones for knee rehabilitation [38], cerebral dysfunction [39] or supporting the education of nutrition [37]. The serious game created in this study falls into the subcategory of digital-game-based learning, which seeks to enhance knowledge and skills through challenges and corresponding accomplishments [40,41].

1.5. JudgED: Serious Game

To provide a possible solution to the above-mentioned issues, a serious game, named JudgED [42] was developed. It is a specialized serious game explicitly designed for Martial Arts referees and judges. It has two main components: an admin interface and a training interface. In the admin interface, the head referee can upload, evaluate, and prepare video clips depicting various fighting scenarios. This interface offers many features, including video editing capabilities like cutting and sequencing, slow motion playback, blurring of judges within the video, and highlighting. Additionally, it allows for the inclusion of metadata for categorization purposes. The head referee can also organize these video clips into courses and playlists for trainees.

In the training interface, another referee can watch and assess the videos regularly, receiving immediate feedback on their judgments. The game encompasses six disciplines, each with its own rule settings, providing different challenges for the players. For instance, in Pointfighting, where only the initial regular technique is awarded a point, the judges stand in the fighting area and use visible hand signs to make decisions. These signs are blurred in the game to prevent the trainee from viewing the live decision. On the other hand, in all other continuous disciplines, the fight continues without interruption after each point. To handle consecutive decisions in these disciplines, the game employs a complex "matching algorithm" [42].

Upon completing a game session, the trainee receives a statistical performance summary, including decision accuracy and the elapsed time between technique execution and decision-making. The home screen also provides overall statistics for multiple game trials. The head referees' home screen features a similar overview but encompasses data from all trainees. In summary, the game was developed based on generalized psychological principles, such as the social cognitive model [7], the multiple-cue probability learning [43], and the Hogarth's approach [44]. These modes are also in (partial) accordance with serious games for other sports, such as soccer [24,25,45], Australian Football [30], or Rugby [8,31].

It should be noted that in contrast to non-digital games, these are often associated with addictive behavior [46]. However, it has also been shown that serious games can increase perceptual and coordination skills through their psychological aspects [46].

1.6. Aims

While JudgED constitutes a supporting system to professionalize the sport further, it has not yet been evaluated. Therefore, this study aims to assess the effectiveness of the JudgED game by (a) measuring decision accuracy and specific reaction time, (b) calculating a theoretical probability of correct point awarding, (c) evaluating the testing reliability, and (d) comparing these results with real-world competition judge agreement data.

1.7. Research Questions and Hypotheses

Formulated in alignment with the overarching objectives, the following subsequent research questions were defined:

- Research Question 1 (RQ1): Can the JudgED system's testing phases be considered reliable?
- Research Question 2 (RQ2): To what extent does the decision accuracy of the JudgED system tests align with the decision accuracy observed in actual, real-life scenarios?

In pursuit of these inquiries, the following hypotheses were subjected to empirical analysis:

Hypothesis 1 (H1). *There are no significant differences between the test and the re-test, using the JudgED system.*

Hypothesis 2 (H2). *There is a significant interrater-reliability agreement between the judges, during the gameplay.*

Hypothesis 3 (H3). *There is a significant interrater-reliability agreement between the judges, judging real tournaments.*

2. Materials and Methods

The following sections describe the used materials and methods for acquiring the results.

2.1. Video Acquisition and Point Definition

The process of video selection, video scene extraction, and definition of decisions was conducted by three subject matter experts: (i) an official WAKO kickboxing referee (experience > ten years, licenses: National A and International Gold Card A), (ii) a former professional WAKO kickboxing athlete (experience 13 years), and (iii) a former professional WAKO kickboxing athlete (experience 16 years) who also has experience as a coach (13 years) and scientist (7 years). The video scenes were only annotated with decisions in case the respective action was visible. Furthermore, to avoid influencing participants of the field experiment by gesticulating referees visible in the video scenes, appearing referees were occluded for the period of the revealing gesture. The judgment difficulty of each video scene was rated by an expert referee, not participating in the experiment, based on a five-point scale covering values from *very low* (1) to *very high* (5). All video scenes (Table 1), including their defined decisions, were additionally reviewed and approved by an international WAKO referee, who was part of the content creation team. The entire content creation and organization process was performed exclusively in JudgED through functionalities of the content and administration module. The video scenes consisted of kickboxing footage from real-world competitions produced by Brannmanndan (https://www.brannmanndan.com, accessed on 1 March 2021). The originator granted permission to use these videos. All videos had a minimum resolution of 720p.

Table 1. Main characteristics of the video scenes for Tatami (V_{T}) and Ring (V_{R}) disciplines, by playlists for the test situations.

Playlist	Number of Video Scenes	Duration	Male/Female	Mean Difficulty
V_{T1}	25	5:25 m	24/1	2.6
V_{T2}	27	6:32 m	23/4	2.6
V_{R1}	12	11:56 m	11/1	3.1
V_{R2}	12	12:40 m	6/6	3.1

The sets V_{T1} and V_{T2} included video scenes of the disciplines Point fighting, Light contact and Kick light. The sets V_{R1} and V_{R2} included video scenes of the disciplines Full contact, Low kick and K1 Style.

2.2. Participants

The sample consisted of 16 licensed (16 for tatami disciplines; 6 additionally for ring disciplines) WAKO kickboxing referees, 2 (12.5%) female and 14 (87.5%) male, with a mean age of 46.6 (±10.8) years and an average refereeing experience of 13.5 (±10.2) years. In order to provide them with appropriate material throughout the field experiment, the sample was divided into the groups S_T (n = 10, age = 46.6 ± 11.5 years, experience = 12.9 ± 9.1 years) and S_R (n = 6, age = 46.7 ± 9.8 years, experience = 14.3 ± 11.5 years).

All participants volunteered to take part in the study and confirmed their agreement by providing written consent. The university's ethical committee granted ethical approval for the study.

2.3. Field Experiment Procedure

The field experiment was divided into two days, including a familiarisation session to reduce bias. The first day involved playing the first experimental playlist, while the second day involved playing the second experimental playlist to evaluate reliability.

Day 1:

1. Familiarization: Prior to the conduction of the subsequent tests, the participants JudgED the set of video scenes V_F to accustom themselves to the functionality of the serious game.
2. Test 1: While the Tatami Referees S_T JudgED video playlist V_{T1}, the Ring referees S_R JudgED playlist V_{R1} in the serious game.

Day 2:

1. Test 2: While the Tatami Referees S_T JudgED playlist V_{T2}, the Ring referees S_R JudgED the videos V_{R2} in the serious game.

All activities were performed on 10-inch Android-based tablets.

2.4. Data Processing

The field experiment will assess the decision accuracy and reaction time of the participating referees. The serious game recorded the correctness and reaction time of each decision (the time between the exact action and the click of the judge playing the game) in the database. The reaction time is available only for decisions that can be matched with defined decisions.

2.5. Qualitative Analysis of WAKO World Championships 2021

As no appropriate literature was found for comparing this particular case, a qualitative analysis was performed on live stream videos of the WAKO World Championships 2021. The overall bout scores, which are the sum of points awarded to each fighter by the judges, were extracted and processed for a total of 476 fights comprising 1399 bouts in continuous disciplines. To narrow down the analysis, we included only the 400 fights with 1200 bouts (involving two fighters, resulting in 2400 decisions) that ended with a winner determined by points (excluding knockouts or technical knockouts) for the statistical procedures.

2.6. Statistical Analysis

All collected data were analysed using IBM SPSS Statistics [47]. Shapiro–Wilk test was performed to test for the normality of the data. All inferential statistic tests were performed at significance level $\alpha = 0.05$.

2.6.1. JudgED Experiment

The Fleiss' kappa inter-rater reliability test was used to measure the degree of agreement among the referees on the judgment of each defined decision and for continuous disciplines of the outcome of every round. To test the data for differences between test 1 and test 2, a t-test for dependent groups was performed. Finally, as a general measure for judging quality with JudgED, the probability for "the correct point awarding" was calculated using the binomial distribution. Therefore, the average decision accuracy of all possible groups of three judges was calculated, and maximum, average, and minimum were used for probability analysis for all disciplines.

$$P = 3x^2 \times (1 - x) + x^3 \tag{1}$$

P = Probability of correct final point awarding x = decision accuracy (per discipline).

2.6.2. Tournament Analysis

The data from the qualitative analysis of the real-world fight analysis were processed on bout level and split in "points for fight winner" and "points for fight loser" for all three judges, resulting in 6 (3 judges × 2 fighters) variables and 7200 data points. The Fleiss' kappa inter-rater reliability test was performed for different gradations, namely (a) each bout and fighter (2400 samples), (b) for each fight and fighter (800 samples), and the overall outcome of each fight (1 = judge awarded win to fight winner; 0 = judge awarded win to fight loser) to measure the degree of agreement among the referees on the judgment of each defined bout. It is noted that Fleiss' kappa considers possible coincidences, which is believed to provide a more accurate measure in this setting than other procedures.

3. Results

The results are split into Section 3.1 including all analyses related to the game and Section 3.2 consisting of the agreement examination between the referees during the competition.

3.1. JudgED System Test

The JudgED game section shares the results obtained from the field experiment's descriptive analysis (Sections 3.1.1 and 3.1.2) on the recorded data. Next, the interrater reliability for the gameplay was tested (Section 3.1.3), followed by an analysis focusing on the two tests' performance reliability of the field experiment (Section 3.1.4). Subsequently, it examines the distinctions between the two assessments and gauges playlist harmonization by evaluating the accuracy according to difficulty levels (Section 3.1.5), by analysing whether the scene-based complexity of playlists is accurate. Lastly, to provide an overall performance assessment, the likelihood of accurate decisions by three judges was computed using the binomial distribution (Section 3.1.6).

3.1.1. Decision Accuracy

The results of the field experiment showed a mean decision accuracy of 43.011% (minimum = 27.047%, maximum = 61.517%, σ = ±12.898%). A Shapiro–Wilk test showed the normal distribution of the users' decision accuracy, $W(16) = 0.889$, $p = 0.053$. Figure 1 visualises the decision accuracy grouped by discipline, which accents the higher decision accuracy of Tatami disciplines (50.460%) compared to Ring disciplines (30.596%).

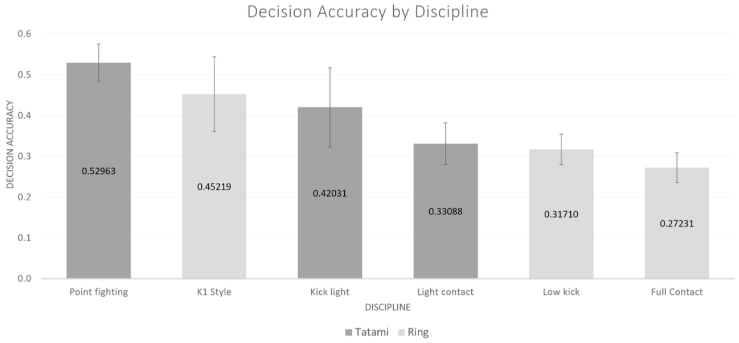

Figure 1. Decision accuracy by discipline for Tatami (dark) and Ring (light), indicating that Point-fighting might be the easiest discipline to judge, due to its point-stop rule, while ring disciplines tend to be more difficult.

3.1.2. Judging Reaction Time

The results of the field experiment showed a mean reaction time of 1.022 s (minimum = 0.755 s, maximum = 1.299 s, $\sigma = \pm 0.156$ s). A Shapiro–Wilk test showed the normal distribution of the users' reaction time, $W(16) = 0.942$, $p = 0.372$. Figure 1 visualises the decision accuracy grouped by discipline, which accents the higher decision accuracy of Tatami disciplines (50.460%) compared to Ring disciplines (30.596%).

The data set was also analysed for the decision value of the defined decision as depicted in Figure 2. While defined decisions for which no user input was expected to be considered correct showed the highest accuracy (57.205%), decisions defined as penalties (i.e., warnings or exits) showed a conspicuously low accuracy (4.475%). Excluding the decisions defined as penalties from the data set would increase the overall decision accuracy from 43.011% to 49.626%.

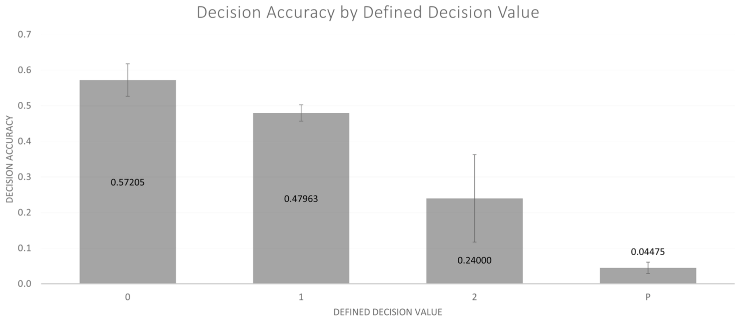

Figure 2. Decision accuracy by defined decision value, indicating that accuracy might correlate with frequency of occurrence.

3.1.3. JudgED Inter-Rater Reliability

The Fleiss' kappa, κ, was calculated to measure the degree of agreement among the referees on the judgment of each defined decision indicated by the combination of decision value and colour. Due to the distinct sets of participants and the different scoring schemes for the Tatami and Ring disciplines, two separate tests were performed.

Ring: The Fleiss' kappa test for Ring was based on the judgment of 297 compelling subjects (defined decisions) from 6 raters (referees). The kappa value of 0.371 indicates a fair strength of agreement between the referees that is statistically significantly different from zero ($p < 0.005$). Table 2 lists the results from the Fleiss' kappa analysis. Further data analysis reveals the differences in agreement between individual rating categories as shown in Table 2. While the kappa coefficients for the decision value *zero* and *one* indicate a good (0.644, 0.633), respectively, fair (0.315, 0.325) level of agreement [48], decisions with value *penalty* show an agreement that lies only slightly above chance agreement (0.024, 0.041). The rating category *no input* refers to defined decisions with values other than zero for which referees did not indicate a judgement. Based on this, an agreement can be assumed as “fair” in both groups.

Table 2. Overall results from the Fleiss’ kappa analysis for Ring and Tatami judges, showing the inter-rater reliability between the judges devided by discipline.

	Kappa	***p***	**n**
Ring	0.371	<0.001	6
Tatami	0.398	<0.001	16

3.1.4. Test 1 and 2 Reliability

Decision accuracy slightly (but not significantly) decreased from 45.333% (±17.194%) to 40.910% (±10.149%), while reaction time increased (not significantly) from 1.008 s (±0.168 s) to 1.036 s (±0.176 s). Further analysis showed that participants of the Tatami disciplines mainly caused a reduction in decision accuracy. While most Ring referees (83.3%) increased their decision accuracy, most Tatami referees (80%) worsened their results from test 1 to test 2.

To examine the differences between test 1 and test 2 of the field experiment, a dependent t-test was computed. A Shapiro–Wilk test was computed to test the assumption for performing the t-test. The differences of the paired values were approximately normally distributed for the measure of decision accuracy, $W(16) = 0.949$, $p = 0.469$, and reaction time,$W(16) = 0.975$, $p = 0.910$. The result of the t-test indicated no significant differences between both tests for decision accuracy, $t(15) = 1.614$, $p = 0.127$, and reaction time, $t(15) = -0.776$, $p = 0.450$.

3.1.5. Examination of Difficulty

Figure 3 visualizes the referees’ performance by the difficulty of video scenes.

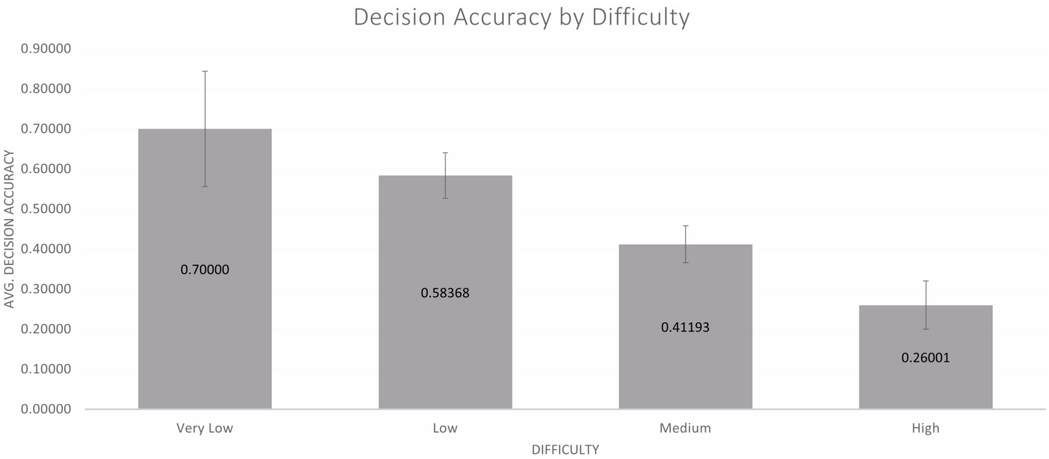

Figure 3. This diagram shows the decision accuracy by difficulty, indicating that the the pre-defined difficulty might be very accurate, which serves as foundation for the playlist selection and ultimately the test–retest comparison.

3.1.6. Probability Analysis

Based on the outcomes, the probability of correct point awarding was calculated using binominal distribution for the three judges with the highest, the lowest, and most average decision accuracy in all disciplines, displayed in Table 3. The lowest probabilities of 15.3% for the lowest ranked three judges in LC and LK show very little chance of a correct final decision. However, it is noted that the low penalty recognition might significantly impact this overall result. It might also heavily affect the highest probability in PF of 67.2%.

Table 3. Binominal Distribution: Probability Analysis for a correct decision of each combination of 3 judges.

	PF	LC	FC	LK	K1
Maximum	67.2%	32.8%	21.2%	25.0%	54.0%
Average	52.8%	24.3%	18.3%	20.0%	44.1%
Minimum	33.3%	15.3%	15.6%	15.3%	32.3%

3.2. Competition Analysis

The Fleiss' kappa, κ, was calculated to measure the degree of agreement among the referees on the judgment of (a) both fighter's final score of each round, (b) both fighter's final fight score, (c) final fight outcome. Results are displayed in Table 4. Based on [48] all results can be classified as slight agreement.

Table 4. Fleiss' kappa analysis for interrater reliability analysis in continuous disciplines at WAKO World Championships, indicating significant low decision reliability.

	Kappa	*p*	n
Round	0.114	<0.001	2400
Fight	0.063	<0.001	800
Outcome	0.166	<0.001	400

4. Discussion

In general, kickboxing judges require training systems that offer comprehensive additional training in situations that closely resemble real-life scenarios. Since the correlation between decision accuracy and the subjective scene difficulty indicates a measure of functionality, the reliability of the test is supported by the absence of significant differences between the first and second tests. Therefore, it can be inferred that the system is suitable for its intended purpose.

4.1. Decision Accuracy and Reaction Time in JudgED

As there were no notable distinctions between both tests regarding decision accuracy and reaction time metrics, hypothesis 1 can be accepted and the first research question answered. The overall findings indicated a mean decision accuracy of 43.011% and a mean reaction time of 1.022 s. However, the low decision accuracy can be partially attributed to the remarkably high error rate observed when assessing penalties (i.e., warnings or exits). Only 4.475% of all penalty decisions were JudgED correctly. By excluding these decisions from the dataset, the overall decision accuracy would increase from 43.011% to 49.626%. Despite the significant interrater reliability, which leads to the acceptance of hypothesis 2, the accordance can be described as somewhere between fair and low. While there exists limited directly analogous literature, the observed outcome of approximately 80% favorable outcomes (indicating erroneous referee decisions) in video reviews within the context of Judo matches suggests a parallel trend [49]. Research in different sports, such as football, however, show a decision accuracy of about 70% [50].

A significant disparity emerged between referees in the Tatami and Ring disciplines after further analysis of decision accuracy. Tatami referees achieved a decision accuracy of 50.460%, whereas Ring referees achieved a lower decision accuracy of 30.596%. This difference can be attributed to the higher average difficulty level of video scenes in the Ring disciplines (3.1) compared to the Tatami disciplines (2.6) and the high number of penalties, which are clearly visible in slow motion, but less so at normal speed. Additionally, video scenes in Ring disciplines contained numerous closely spaced decisions, which are believed to be more challenging to judge. On average, video scenes in the Ring disciplines contained 12.375 decisions, while those in the Tatami disciplines had an average of only 2.288 decisions.

4.1.1. Probability Analysis

The analysis revealed that considering all decisions, the probability for three referees to accurately judge the correct decision ranges from 15.3% to 67.2%. However, it is essential to note that this probability does not directly represent the likelihood of the correct outcome of the fight. It should be acknowledged that certain factors, such as a "missed" penalty, can occur within the data but may not necessarily impact the overall fight outcome if it results in a "verbal warning" or happens once in a fight for example, it does not affect the points at all. Furthermore, multiple decisions are made within a bout, and multiple bouts are conducted within a fight. This means that if one fighter is dominant, the probability of them losing due to judge performance is still much lower than the binomial distribution results suggest. Nonetheless, these statistics still raise concerns.

4.1.2. Competition Analysis

The analysis using Fleiss' kappa method shows that there is a noticeable but not very strong agreement (ranging from 0.063 to 0.166) among the three judges during the fights, which leads to the acceptance of hypothesis 3. The reason for the lower agreement among judges might be because they hold different opinions, even though we expected them to agree more. Interestingly, we observed a slightly better agreement when using the JudgED system, which suggests that it could be useful for training. However, it is important to note that we cannot make a perfect comparison between the two methods because they collect data in different ways.

4.2. JudgED Outcome Comparison with Competition Analysis

It is worth noting that the presented methodology does not permit a perfect comparison. Nevertheless, it provides a limited and rough estimation to answer the question, if the JudgED study outcomes reflect reality. Based on the similarly low agreement within the game and during the competition, it is stated that the outcomes might reflect reality, also answering research question 2. It is important to emphasize that the lower accuracy in the JudgED system test does not reflect problems with the system itself. Instead, it sheds light on the judges' overall performance and the specific difficulties associated with the sport.

4.3. Limitations

The main constraint of this study concerns the experimental design utilized to evaluate the referees' performance in the serious game. Several factors contribute to this limitation: The relatively low number of participants in the experiment may restrict the generalization of the findings. The brief duration given to familiarise oneself with the mechanics of the serious game may have adversely affected performance. Another aspect to consider is that the viewing perspective in the serious game does not precisely represent the natural competitive decision-making environment. The video scenes consist of footage from real-life competitions and are not recorded from a first-person perspective. The field experiment was conducted on 10-inch tablets, which may have posed challenges in accurately identifying certain decisions. It is worth noting that the real-competition and field test data in JudgED differ to some extent. In the serious game, complete bouts are not scored; all judges are exposed to the same scene and perspective. However, finding a suitable, feasible, and fully comparable solution does require further advanced analysis methods and development, which are not technically feasible at present but may be the subject of future evaluation.

5. Outlook

To further improve and evaluate the known data, a larger study will be carried out after this pilot test. This study will comprise the analysis of a championship's Sportdata (official online scoring provider) logfiles, a training intervention with the JudgED system, and another log file analysis on a tournament.

6. Conclusions

It is important to bear in mind that contests are also constrained by time. As a result, it is challenging to acquire expertise if one aims to obtain the required hours or matches. This article focuses exclusively on referees in combat sports. The employment of the serious game introduced herein represents one potential method for skill development.

The outcomes presented in this article demonstrate that the training with JudgED exhibits almost the same decision-making standards as real contests. This would imply that players can be evaluated and trained without the need for participation in tournaments, leading to an enormous improvement in quality. Furthermore, minimizing travel time would conserve time and contribute to environmental protection.

Author Contributions: Conceptualization, D.H., C.G. and T.G.; methodology, D.H., S.Ö., S.U., A.B.O. and C.G.; software, A.S.; validation, D.H., A.S. and C.G.; formal analysis, A.S. and C.G.; investigation, D.H., A.S. and C.G.; resources, S.Ö., S.U. and T.G.; data curation, D.H., A.S. and C.G.; serious gaming expertise and conceptualization, R.B.; writing—original draft preparation, D.H.; writing—review and editing, S.Ö., S.U., A.B.O., R.B., K.P. and T.G.; visualization, A.S. and K.P.; supervision, T.G.; project administration, D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Ethics Committee of Vienna University of Technology.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Raw data are not publicly available.

Acknowledgments: First and foremost we would like to thank Lucian Gutica-Florescu, who contributed to the development of the JudgED game alongside Andre Salmhofer. Furthermore, we would like to thank all WAKO officials and referees who supported the project, helped with content creation and participated in this study, as well as Brannmanndan for access to the videos.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

JudgED	Judge Education Serious Game
WAKO	World Association of Kickboxing Organisations
PF	Pointfighting (intermitted tatami discipline)
LC	Light Contact (continuous tatami discipline)
KL	Kick Light (continuous tatami discipline)
FC	Full Contact (continuous ring discipline)
LK	Low Kick (continuous ring discipline)
K1	K1 (continuous ring discipline)

References

1. MacMahon, C.; Strauss, B. The psychology of decision making in sports officials. In *Routledge Companion to Sport and Exercise Psychology*; Routledge: Milton, UK, 2014; pp. 223–235.
2. Williams, A.M.; Jackson, R.C. *Anticipation and Decision Making in Sport*; Routledge: Milton, UK, 2019.
3. Hölbling, D.; Baca, A.; Dabnichki, P. A kinematic model for assessment of hip joint range-of-motion in fast sport movements using spreading angles. *Sport. Biomech.* **2020**, 1–13. [CrossRef] [PubMed]
4. Hoelbling, D.; Baca, A.; Dabnichki, P. Sequential action, power generation and balance characteristics of a martial arts kick combination. *Int. J. Perform. Anal. Sport* **2020**, *20*, 766–781. [CrossRef]
5. Carlsson, T.; Berglez, J.; Koivisto Persson, S.; Carlsson, M. The impact of video review in karate kumite during a Premier League competition. *Int. J. Perform. Anal. Sport* **2020**, *20*, 846–856. [CrossRef]
6. Helsen, W.F.; MacMahon, C.; Spitz, J. Decision making in match officials and judges. In *Anticipation and Decision Making in Sport*; Routledge: Milton, UK, 2019; pp. 250–266.
7. Kittel, A.; Cunningham, I.; Larkin, P.; Hawkey, M.; Rix Lievre, G. Decision-making training in sporting officials: Past, present and future. *Psychol. Sport Exerc.* **2021**, *56*, 102003. [CrossRef]

8. Mascarenhas, D.; O'Hare, D.; Plessner, H. The psychological and performance demands of association football refereeing. *Int. J. Sport Psychol.* **2006**, *37*, 99–120.
9. Moore, B.; Dudley, D.; Woodcock, S. The effect of martial arts training on mental health outcomes: A systematic review and meta-analysis. *J. Bodyw. Mov. Ther.* **2020**, *24*, 402–412. [CrossRef]
10. Jhon, A. The influence of playing 3D fighting games in practicing martial arts. *J. Phys. Conf. Ser.* **2019**, *1175*, 12253. [CrossRef]
11. World Association of Kickboxing Organizations (WAKO). WAKO Kickboxing Rules. 2022. Available online: https://wako.sport/wp-content/uploads/2022/10/WAKO-Rules-25.10.2022.-revision-3.pdf (accessed on 15 May 2023).
12. Ouergui, I.; Hssin, N.; Haddad, M.; Franchini, E.; Behm, D.G.; Wong, D.P.; Gmada, N.; Bouhlel, E. Time-motion analysis of elite male kickboxing competition. *J. Strength Cond. Res.* **2014**, *28*, 3537–3543. [CrossRef]
13. Slimani, M.; Chaabene, H.; Miarka, B.; Chamari, K. The activity profile of elite low-kick kickboxing competition. *Int. J. Sport. Physiol. Perform.* **2017**, *12*, 182–189. [CrossRef] [PubMed]
14. Ouergui, I.; Hssin, N.; Franchini, E.; Gmada, N.; Bouhlel, E. Technical and tactical analysis of high level kickboxing matches. *Int. J. Perform. Anal. Sport* **2013**, *13*, 294–309. [CrossRef]
15. Bless, H.; Fiedler, K.; Strack, F. *Social Cognition: How Individuals Construct Social Reality*; Psychology Press: London, UK, 2004.
16. Silva, A.F.; Conte, D.; Clemente, F.M. Decision-making in youth team-sports players: A systematic review. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3803. [CrossRef]
17. Edwards, W. The theory of decision making. *Psychol. Bull.* **1954**, *51*, 380. [CrossRef] [PubMed]
18. Calabretta, G.; Gemser, G.; Wijnberg, N.M. The Interplay between Intuition and Rationality in Strategic Decision Making: A Paradox Perspective. *Organ. Stud.* **2017**, *38*, 365–401. [CrossRef]
19. Daniel, K. *Thinking, Fast and Slow*; Penguin: London, UK, 2012.
20. Araujo, T.; Helberger, N.; Kruikemeier, S.; de Vreese, C.H. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI Soc.* **2020**, *35*, 611–623. [CrossRef]
21. Curley, L.J.; Munro, J.; Dror, I.E. Cognitive and human factors in legal layperson decision making: Sources of bias in juror decision making. *Med. Sci. Law* **2022**, *62*, 206–215. [CrossRef]
22. Anwar, S.; Bayer, P.; Hjalmarsson, R. Politics in the Courtroom: Political Ideology and Jury Decision Making. *J. Eur. Econ. Assoc.* **2018**, *17*, 834–875. [CrossRef]
23. Petersen, T.S.; Wichmann, S.S. Fairness, implicit bias testing and sports refereeing: An argument for why professional sports organisations ought to promote fairness in sport through testing referees for implicit biases. *J. Philos. Sport* **2021**, *48*, 97–110. [CrossRef]
24. Plessner, H.; Haar, T. Sports performance judgments from a social cognitive perspective. *Psychol. Sport Exerc.* **2006**, *7*, 555–575. [CrossRef]
25. Schweizer, G.; Plessner, H.; Kahlert, D.; Brand, R. A Video-Based Training Method for Improving Soccer Referees' Intuitive Decision-Making Skills. *J. Appl. Sport Psychol.* **2011**, *23*, 429–442. [CrossRef]
26. Brunswik, E. *The Conceptual Framework of Psychology*; University Chicago Press: Chicago, IL, USA, 1952.
27. Larkin, P.; Berry, J.; Dawson, B.; Lay, B. Perceptual and decision-making skills of Australian football umpires. *Int. J. Perform. Anal. Sport* **2011**, *11*, 427–437. [CrossRef]
28. Macmahon, C.; Helsen, W.; Starkes, J.; Weston, M. Decision—making skills and deliberate practice in elite association football referees. *J. Sport. Sci.* **2007**, *25*, 65–78. [CrossRef]
29. Ericsson, K.; Krampe, R.; Tesch-Roemer, C. The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychol. Rev.* **1993**, *100*, 363–406. [CrossRef]
30. Larkin, P.; Mesagno, C.; Berry, J.; Spittle, M.; Harvey, J. Video-based training to improve perceptual-cognitive decision-making performance of Australian football umpires. *J. Sport. Sci.* **2018**, *36*, 239–246. [CrossRef]
31. Mascarenhas, D.; Collins, D.; Mortimer, R.; Morris, B. A naturalistic approach to training accurate and coherent decision making in Rugby Union Referees. *Sport Psychol.* **2005**, *19*, 131–147. [CrossRef]
32. Put, K.; Wagemans, J.; Spitz, J.; Williams, A.M.; Helsen, W.F. Using web-based training to enhance perceptual-cognitive skills in complex dynamic offside events. *J. Sport. Sci.* **2016**, *34*, 181–189. [CrossRef] [PubMed]
33. Cizmic, D.; Hoelbling, D.; Baranyi, R.; Breiteneder, R.; Grechenig, T. Smart Boxing Glove "RD α": IMU Combined with Force Sensor for Highly Accurate Technique and Target Recognition Using Machine Learning. *Appl. Sci.* **2023**, *13*, 9073. [CrossRef]
34. Hoelbling, D.; Grafinger, M.; Smiech, M.M.; Cizmic, D.; Dabnichki, P.; Baca, A. Acute response on general and sport specific hip joint flexibility to training with novel sport device. *Sport. Biomech.* **2021**, 1–16. [CrossRef]
35. Michael, D.R.; Chen, S.L. *Serious Games: Games That Educate, Train, and Inform*; Muska & Lipman/Premier-Trade: New York, NY, USA, 2005.
36. Baranyi, R.; Czech, P.; Walcher, F.; Aigner, C.; Grechenig, T. Reha@Stroke—A Mobile Application to Support People Suffering from a Stroke Through Their Rehabilitation. In Proceedings of the IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH), Kyoto, Japan, 5–7 August 2019; pp. 1–8.
37. Aigner, C.; Resch, E.M.; El Agrod, A.; Baranyi, R.; Grechenig, T. Food Pyramid Escape-A serious escape game for the support of nutritional education in Austria and beyond. In Proceedings of the IEEE 9th International Conference on Serious Games and Applications for Health (SeGAH), Dubai, United Arab Emirates, 4–6 August 2021; pp. 1–8.

38. Rast, L.; Baranyi, R.; Pinter, K.; Hölbling, D.; Aigner, C.; Grechenig, T. Standard Mobile Phones Plus a Balance Board Are Sufficient: Designing a Serious Game for Better Knee Rehabilitation. In *dHealth 2023*; IOS Press: Amsterdam, The Netherlands, 2023; pp. 18–19.
39. Baranyi, R.; Perndorfer, R.; Lederer, N.; Scholz, B.; Grechenig, T. MyDailyRoutine-a serious game to support people suffering from a cerebral dysfunction. In Proceedings of the IEEE International Conference on Serious Games and Applications for Health (SeGAH), Orlando, FL, USA, 11–13 May 2016; pp. 1–6.
40. Qian, M.; Clark, K.R. Game-based Learning and 21st century skills: A review of recent research. *Comput. Hum. Behav.* **2016**, *63*, 50–58. [CrossRef]
41. Pagé, C.; Bernier, P.M.; Trempe, M. Using video simulations and virtual reality to improve decision-making skills in basketball. *J. Sport. Sci.* **2019**, *37*, 2403–2410. [CrossRef]
42. Salmhofer, A.; Gutica-Florescu, L.; Hoelbling, D.; Breiteneder, R.; Baranyi, R.; Grechenig, T. Development of a Serious Game to Improve Decision-making Skills of Martial Arts Referees. In Proceedings of the 10th International Conference on Sport Sciences Research and Technology Support, Valletta, Malta, 27–28 October 2022.
43. Lagnado, D.A.; Newell, B.R.; Kahan, S.; Shanks, D.R. Insight and strategy in multiple-cue learning. *J. Exp. Psychol. Gen.* **2006**, *135*, 162–183. [CrossRef]
44. Hogarth, R.M. On the learning of intuition. In *Intuition in Judgment and Decision Making*; Psychology Press: London, UK, 2011; pp. 111–126.
45. Gulec, U.; Yilmaz, M. A serious game for improving the decision making skills and knowledge levels of Turkish football referees according to the laws of the game. *SpringerPlus* **2016**, *5*, 1–10. [CrossRef]
46. Wiemeyer, J. Gesundheit auf dem Spiel?—Serious Games in Prävention und Rehabilitation. *Dtsch. Z. Für Sportmed.* **2010**, *61*, 252–257.
47. Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.R.; Wirth, R. *CRISP-DM 1.0: Step-by-Step Data Mining Guide*; SPSS Inc.: Chicago, IL, USA, 2000.
48. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159. [CrossRef] [PubMed]
49. Kons, R.L.; Saldanha Da Silva Athayde, M.; Ceylan, B.; Franchini, E.; Detanico, D. Analysis of video review during official judo matches: Effects on referee's decision and match results. *Int. J. Perform. Anal. Sport* **2021**, *21*, 555–563. [CrossRef]
50. Kittel, A.; Larkin, P.; Elsworthy, N.; Lindsay, R.; Spittle, M. Effectiveness of 360 virtual reality and match broadcast video to improve decision-making skill. *Sci. Med. Footb.* **2020**, *4*, 255–262. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.