TU WIEN Informatics

Design, Development and Evaluation of a Serious Game to Train Decision-making Skills of Martial Arts Referees

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Software Engineering and Internet Computing

eingereicht von

André Salmhofer, BSc.
Matrikelnummer 01025963

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Thomas Grechenig
Mitwirkung: Dr. Dominik Hölbling, MSc, Bakk.

Wien, 26. Juni 2023

Unterschrift Verfasser

Unterschrift Betreuung

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

TU WIEN Informatics

Design, Development and Evaluation of a Serious Game to Train Decision-making Skills of Martial Arts Referees

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Software Engineering and Internet Computing

by

André Salmhofer, BSc.
Registration Number 01025963

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Thomas Grechenig
Assistance: Dr. Dominik Hölbling, MSc, Bakk.

Vienna, 26th June, 2023

Signature Author

Signature Advisor

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Design, Development and Evaluation of a Serious Game to Train Decision-making Skills of Martial Arts Referees

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Software Engineering and Internet Computing

eingereicht von

André Salmhofer, BSc.
Matrikelnummer 01025963

ausgeführt am
Institut für Information Systems Engineering
Forschungsbereich Business Informatics
Forschungsgruppe Industrielle Software
der Fakultät für Informatik der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Thomas Grechenig

Wien, 26. Juni 2023

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

André Salmhofer, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 26. Juni 2023

André Salmhofer

Acknowledgements

First and foremost, I would like to thank my mother Evelyn, my partner Anna, and my friends for supporting me throughout my academic career. Many thanks also go to my thesis advisors Thomas Grechenig and Dominik Hoelbling, head and member of the research group for industrial software, for their continuous support. In addition to his expertise in scientific topics and kickboxing, his network made it possible to get in touch with a large number of subject matter experts who provided crucial input for the design and development of the prototype. Furthermore, I would like to thank *World Association of Kickboxing Organizations* (WAKO), which contributed to the project's success by providing their support during the design, development, and evaluation of the prototype. This allowed for the involvement of experts, the promotion of the project within the WAKO kickboxing referee community and the evaluation of the developed prototype with elite kickboxing referees. I would particularly like to thank Günter Ecker, an experienced WAKO reviewer who significantly contributed to the success of the research in many aspects. In addition to sharing his expertise in the steps of analysis, design, and development, he also prepared most of the video scenes in JudgED and served as a facilitator to organise the evaluation of the prototype during the Austrian WAKO Kickboxing Championship 2022 in Graz. I also would like to thank Daniel Slevin, the owner and producer of the kickboxing video channel *Brannmanndan*, for permission to use his videos in JudgED. Finally, I would like to thank Roland Breiteneder, CEO of Sportdata GmbH, for his support, particularly in providing the company's Vimeo account required for hosting the video content used in the developed prototype.

Kurzfassung

Während Schiedsrichter ein breites Spektrum hinsichtlich Wahrnehmung, Fitness und Interaktion mit Athleten abdecken, liegt ihr Fokus im Treffen von Entscheidungen, um die Regeln der Sportart durchzusetzen. Da Kampfsport-Athleten dicht aufeinanderfolgende Techniken ausführen können, müssen Kampfsport-Schiedsrichter in der Lage sein die visuelle Wahrnehmung mit den Regeln der Sportart zu kombinieren, um eine schnelle Entscheidung abzuleiten. Um die für das Erreichen von Expertenniveau notwendige Trainingsintensität anzusammeln, schlagen neue Forschungsergebnisse den Einsatz videobasierter Trainingsprogramme vor. Die Integration von Spielelementen in ein solches Programm kann den Lernprozess in ein intrinsisch motiviertes Erlebnis verwandeln.

Diese Arbeit umfasst den Entwurf, die Entwicklung, und die Evaluierung eines videobasierten Serious Games, um Entscheidungsprozesse von Kampfsportschiedsrichtern mittels direkten Feedbacks zu trainieren. Der Prototyp *JudgED* wurde nach Prinzipien des Multiple-Cue Probability Learning und Hogarth's Ansatz zum Lernen von Intuitionen entwickelt. Das Programm basiert auf Erkenntnissen der Forschung zu videobasierten Entscheidungsprogrammen und Inputs von Experten in Kickboxen und Karate Kumite.

JudgED besteht aus zwei Modulen: (a) einem Trainingsmodul, um die Entscheidungen von Kampfsport-Schiedsrichtern durch einen videobasierten Ansatz mit sofortigem Feedback zu trainieren und (b) einem Inhalts- und Administrationsmodul zum Definieren und Organisieren der Videoszenen, sowie zur Leistungsanalyse der Spieler des Serious Games.

JudgED wurde im Zuge eines Feldexperiments evaluiert, bei welchem die Leistung von 16 professionellen Kickbox-Schiedsrichtern analysiert wurde. Das Experiment bestand aus zwei in JudgED durchgeführten videobasierten Tests und einem allgemeinen Reaktionszeittest. Die Leistungsdaten wurden durch ein klar definiertes Verfahren berechnet, welches die Eingaben der Teilnehmer mit von Experten definierten Entscheidungen vergleicht. Die Ergebnisse zeigen eine Entscheidungsgenauigkeit von 43,011 % und eine Reaktionszeit von 1,022 s. Es wurde kein Zusammenhang zwischen der Reaktionszeit in JudgeED und dem Reaktionszeittests gefunden. Ebenso wurde keine Korrelation zwischen der Erfahrung von Schiedsrichtern und deren Leistung in JudgED festgestellt. Die Ergebnisse einer Umfrage zeigen den potenziellen Nutzen von JudgED zum Training von Kampfsport-Schiedsrichtern und zur Aufwertung der Schiedsrichterausbildung.

Keywords: *Entscheidungstraining, Serious Game, Digital game-based learning, Schiedsrichter, Kampfsport*

Abstract

While referees must cover a wide range of demands related to perception, physical fitness and interaction with athletes, the distinctive responsibility associated with referees lies in decision-making to enforce the rules of the sport. As athletes can perform multiple rapid techniques in a short period, martial arts referees focus on deriving decisions by combining their perception of the athletes' movement with the rules of the sport. To achieve the training intensity required to reach expert level in decision-making, recent research suggests the complementary use of video-based training programs. The integration of game elements into video-based training programs has the potential to turn learning into a joyful and intrinsically motivated experience.

This research encompasses the design, development, and evaluation of a novel video-based serious game to train the decision-making processes of martial arts referees through immediate feedback. The prototype *JudgED*, was designed according to the principles of multiple-cue probability learning and Hogarth's approach to learning intuitions. It was developed based on state-of-the-art research on video-based decision-making programs in other sports and requirements gathered from experts in kickboxing and karate Kumite.

JudgED consists of two modules: (a) a training module to train decision-making skills of martial arts referees through a video-based approach utilising immediate feedback and (b) a content and administration module allowing authorised referees to define video scenes used in the serious game, organise training sessions, and analyse the players' performance.

JudgED was evaluated in the form of a field experiment that analysed the performance of 16 professional kickboxing referees in the serious game. The field experiment consisted of two video-based tests in the serious game and a general reaction time test. The performance data in JudgED was determined by a well-defined procedure that compares the players' inputs in the serious game with expert-defined decisions. The findings showed an average decision accuracy of 43.011 % and an average reaction time of 1.022 s. No significant linear relationship was found between referees' reaction time in JudgeED and their performance in the reaction time tests. Likewise, the data show no significant correlation between the referees' experience and their performance in JudgED. The results of a survey indicate the usefulness of JudgED for training the decision-making skills of martial arts referees and its potential to enhance referee training.

Keywords: *Decision-making training, Serious game, Digital game-based learning, Referees, Judges, Martial arts.*

Contents

CHAPTER 1

Introduction

1.1 Problem Statement

While athletes are the centre of attention in sports competitions, referees are often perceived as neglected participants performing a straightforward task [MS14]. However, scrutinising the responsibilities and activities a referee needs to master throughout sports competitions reveals its difficulty and emphasises the importance of referees as crucial stakeholders along with players, coaches, and spectators [KCL+21].

Referees need to cover a broad spectrum of skills in perception, physical fitness and interaction with athletes depending on the characteristics of the sport. While the emphasis of these demands varies among different types of referees, the characteristic responsibility referees are connoted with is the task of judgement and decision-making to enforce the rules of the sport as a prerequisite to determine the winner following a fair competition [MS14].

Examining the decision-making task in detail reveals it to be a complex social-cognitive process influenced by various external constraints specific to the officiated sport [KCL+21]. It requires a combination of declarative knowledge comprising the rules of the sport and procedural knowledge obtained through practical experience [LMB+18]. Due to the reason that athletes can execute multiple fast-moving techniques within a short period, the challenge of martial arts referees is to derive a proper judgement from memory by combining the perception of the athletes' movement with the prior experience and rules of the sport [CBKPC20]. Rather than physical fitness and interaction, the focus lies on perception and cognitive processing [MS14]. If referees are not adequately trained or supported by digital review techniques as partly used in karate Kumite [CBKPC20], football [SWM+21], and tennis [KSR17], the complexity of this process can lead to decision errors [CBKPC20, CGGA+10, OVB+00, CVS16, Mat08], which consequently lead to wrong outcomes of single competitions or even tournaments.

1.2 Motivation

The potential influence on the game's outcome and derived economic consequences caused by decision errors resulted in an increased investigation of officials' decisions [LBDL11]. The literature enumerates several approaches to training sports officials' decision-making skills. Although directly participating in sports competitions is acknowledged as an ideal method to develop decision-making skills [KCL+21, MHSW07], the number of competitive events throughout the year limits its applicability. According to skill development frameworks, like the 10.000h rule of deliberate practice [EKTR93], a high magnitude of training intensity needs to be accumulated to reach expert level in certain skills. For example, assuming an average duration of ten minutes per fight in kickboxing, referees would need to officiate 60.000 fights until reaching expert level in the area of decision-making.

An alternative training modality is conducting simulated drills to train certain elements present in competitions. However, due to the practical and logistical challenges of bringing athletes together and role-playing realistic decision-making scenarios, this approach may not provide the required training intensity to reach the [KCL+21] expert level. Thus, there is a demand to examine complementary training modalities to train decision-making skills without being solely dependent on on-field experience.

A potential solution to address the shortcomings of on-field training modalities is the application of video-based training programs designed to train decision-making skills. The development of theoretically-grounded video-based decision-making training tools emerged over the past 17 years [KCL+21]. Video-based training programs enable referees to reach a high level of practical training intensity, which would hardly be possible by solely judging real-world competitions [LMB+18].

For sports like rugby [MCMM05], Australian football [LMB+18] and soccer [SPKB11, GY16, PWS+16] studies about video-based training programs for referees, evidencing their effectiveness, were already conducted. However, based on the current knowledge, research has yet to examine video-based training programs to train decision-making skills of martial arts referees.

1.3 Aim of the Work

The scope of the thesis was to design and develop a video-based serious game to train decision-making processes of martial arts referees through immediate feedback [SPKB11]. The developed serious game can be further classified as digital game-based learning, which aims to enhance knowledge and skills promoted by challenges and linked achievements [QC16]. The prototypical implementation of the serious game is used to examine the *decision accuracy* and *reaction time* of licensed martial arts referees. It also evaluates if the serious game has the potential to enhance the training of martial arts referees. Thus, this work examines whether the positive results of video-based decision-making training,

evidenced by studies of other sports, can also be transferred to the area of martial arts refereeing.

While this thesis covers the requirements engineering, design, and evaluation for the entire prototype, the implementation-related tasks are reduced to features comprising module ***b*** as well as cross-cutting concerns. In order to establish the full functionality of the prototype, features covering module ***a*** are delivered through cooperation with another student.

Besides the ability to conduct a user study with the developed serious game, it should reach maturity on a level that it can be piloted as a complementary training method in referee education. This will allow further studies to evaluate the acceptance, effectiveness and ability to transfer the gained decision-making skills to real-world competitions.

1.4 Research Questions

This work examines the following research questions (RQ), which cover the areas of analysis, development and evaluation of the developed prototype:

- **RQ_1:** Which requirements can be identified for a serious game to train decision-making skills of martial arts referees in terms of decision accuracy and reaction time?
- **RQ_2:** How can the serious game accurately determine the correctness and reaction time of the judgments entered by the martial arts referee based on the events appearing in the fight scene presented in the form of a streaming video?
- **RQ_3:** How do professional martial arts referees perform in the serious game in terms of decision accuracy and reaction time, and how does this relate to their refereeing experience and general reaction time?
- **RQ_4:** How do professional martial arts referees assess the potential of the serious game to enhance referee training?

In order to further specify the research questions RQ_3, the following hypotheses are made:

- **$H_{3.1}$:** The decision accuracy of referees in the serious game does not improve between two consecutive tests performed in the serious game without feedback.
- **$H_{3.2}$:** The average decision accuracy of referees in the serious game is greater than 50 %.
- **$H_{3.3}$:** The level of refereeing experience has a positive impact on the decision accuracy in the serious game.

- **$H_{3.4}$:** Video scenes with increased difficulty rank show a reduced decision accuracy in the serious game.
- **$H_{3.5}$:** The general reaction time of referees has a positive impact on the reaction time in the serious game.
- **$H_{3.6}$:** The level of agreement among the referees in the serious game on the judgments of expert-defined decisions is *moderate* [KL16].

In order to examine the research questions mentioned above, this work covers the creation of various artefacts. First, a requirements catalogue is created, which provides the basis for designing the serious game. In addition, mockups are drafted, which serve as an instrument to converge to the final design of the serious game's user interface fulfilling the stakeholders' expectations and closing requirements gaps. Based on the two artefacts mentioned above, an executable prototype is developed and documented in technical and functional aspects. In addition, a questionnaire is designed to collect feedback on the developed prototype.

Depending on the research question, different methods are applied to answer them properly. Section 3 describes the methods used to answer each research question.

CHAPTER 2

State of the Art

In order to place the research on a solid knowledge base, this chapter describes the theoretical foundations of this thesis. It covers the topics decision making of referees (section 2.1), serious games (section 2.2), prototyping (section 2.4), components of web-based systems (section 2.5), questionnaire design (section 2.6), statistical methods (section 2.7), and related work (section 2.8).

2.1 Referees & Decision-Making

While athletes and coaches are acknowledged as key stakeholders in sports competitions, referees are often perceived as neglected participants performing a straightforward task. As referees' decisions can potentially influence the outcome of competitions, they often cause anger and frustration among the spectators. Due to the technical possibilities to instantly analyse critical situations, unambiguously wrong referee decisions can be quickly revealed, which causes criticism and leads to questioning the skills of referees. While bad decisions are eye-catching, excellent decisions often remain unrecognised even if the respective judgment prevented a potential dispute. Thus, remaining unnoticed is an indication of a good refereeing performance in many sports [MS14]. Scrutinising the tasks a referee needs to manage, discloses its difficulty and underlines the importance of referees as key stakeholders along with players, coaches and spectators [KCL+21].

This section provides an overview of the domain of refereeing, starting with a classification of referees determined by the characteristics and challenges associated with the judged sport. Then, to grasp the complexity of making appropriate decisions, the decision-making process is examined in detail by considering various influencing factors.

2.1.1 Definition & Classification of Referees

Examining the role of the official discloses its diversity with respect to the variety of the demanded abilities. For example, when comparing soccer referees with gymnastic judges, a considerable difference becomes apparent. Although both roles fall under the term of an official, their emphases are very different [Mac15, p. 1]. According to the Australian Department of Local Government, Sport and Cultural Industries [DoLGSR], the role of the official is abstractly defined as follows:

> *"An official is any person who controls the actual play of a competition by the application of the rules and laws of the sport to make judgements on the rule infringements, performance, time or score."*

To refine this definition, it is helpful to differentiate between different types of officials. Depending on the characteristics of the judged sport, referees need to cover a broad spectrum of skills in perception, physical fitness and interaction with athletes [MS14]. Therefore, referees can be classified into different types based on (i) the number of cues or athletes to be monitored and (ii) the degree of interaction with athletes and physical movement demands. Derived from these classifying dimensions, three types of officials can be distinguished: *interactors*, *monitors* and *reactors* [Mac15, p. 9]. Figure 2.1 visualises the boundaries between the three different referee types according to their classifying dimensions, together with examples of referees in different sports.

Interactors are associated with high interaction and movement demands, often requiring dealing with many athletes or cues. This referee type covers a broad range of officials like soccer, basketball and wrestling referees [Mac15, p. 9]. In order to keep up with the play and ensure an optimal viewing position, interactors often need to cover large distances. For example, while basketball referees cover a distance of four to six kilometres per match, soccer referees cover an average distance of ten kilometres in different paces ranging from walking to sprinting, which is comparable to a midfield player [Mac15, p. 33-36].

Monitors, such as volleyball referees and gymnastic judges, are characterised by low to medium interaction and movement demands. The complexity of their task is determined by the necessity to track and process a medium to high number of athletes or cues [Mac15, p. 9].

Reactors are characterised by low interaction and movement demands as well as a low to medium number of cues and athletes to be processed. An example of this category are tennis line judges [Mac15, p. 9].

Referring to these definitions, Carlsson et al. [CBKPC20] classify karate Kumite judges as monitors, underpinned by the argument that they are faced with the same challenge as gymnastic judges in terms of assessing multiple consecutive techniques within a short period. Thus, the challenge is to derive a proper judgement from memory by combining the perception of the athletes' movement with the prior experience and rules of the sport.

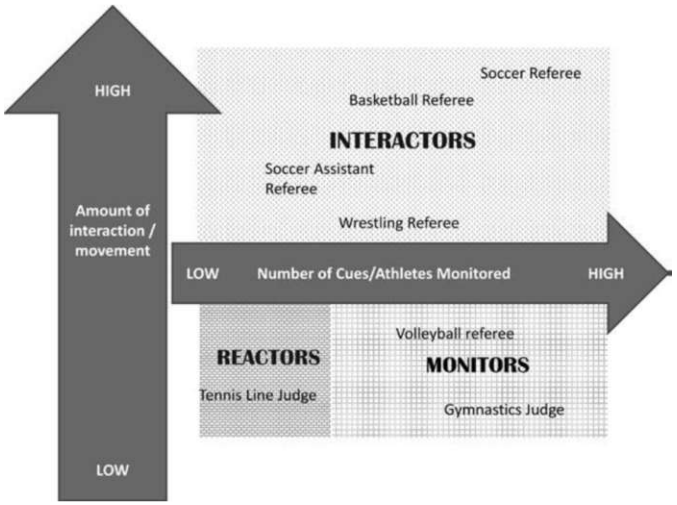


Figure 2.1: Classification of officials [Mac15, p. 10].

Rather than physical fitness and interaction, the focus lies on perception and cognitive processing [MS14]. Considering the similarities between judging fights in karate Kumite and kickboxing, kickboxing referees might also be classified as monitors. In many sports, referees and judges work together as a team to derive appropriate decisions. For example, certain decisions in professional WAKO kickboxing competitions are derived based on a majority decision among three independent judges [waka].

2.1.2 The Role of Decision-Making in Sport

The previous section provided an overview of the main demands of various types of officials. While the required skills vary among the different types of referees, the characteristic responsibility referees are associated with is the task of judgement and decision-making to enforce the rules of the sport and ensure a fair competition [MS14]. Examining the decision-making task in detail discloses a complex social-cognitive process influenced by various external constraints specific to the officiated sport [KCL+21]. It requires a combination of declarative knowledge comprising the rules of the sport and procedural knowledge obtained through practical experience [LMB+18].

Stefani and Bennett [SB98] introduce a taxonomy that classifies sports with respect to their type and mode of evaluation. The taxonomy suggests that performances in sports competitions can be evaluated by objective measurements (e.g. in swimming or weightlifting), objective scores (e.g. in soccer or golf) or performance judgments (e.g. in combat sports, diving, gymnastics). The importance of judgments as an essential part of competitive sports is underpinned by the fact that almost one-third of all sports recognised by the International Olympic Committee uses a performance rating system to evaluate the performance of athletes. Except for epee fencing, which is scored electronically, all combat sports are evaluated based on performance judgements. Even in sports predominantly

assessed by objective measures or scores, judgements of critical situations are inevitable, such as the judgment of tackling situations in soccer [PH06].

Judgments of sports performances can be further divided into three dimensions: (i) evaluative judgements, (ii) judgments of identification, and (iii) judgments of cause. While evaluative judgments assess performances based on a good-bad scale, judgments of identification are concerned with recognising certain situations according to the sport's rules. Judgments of cause are concerned with the determination of causes contributing to specific outcomes. The subsequent paragraphs exclusively focus on evaluative judgments and judgments of identification in sports, which are the subject of most studies examining referee decision-making [PH06].

2.1.3 Referee Decision-Making Process and Influencing Factors

Social cognition research can be used to understand the single steps involved in the decision-making processes of officials, which investigates the social knowledge and cognitive processes involved when people construct their subjective reality [FT91, PH06]. Following an information processing framework, social cognition research examines how individuals make judgments, attributions and decisions [BFS04]. The applicability of this process to describe the judgment of sports performances is justified by the assumption that it follows the general principles of social judgments [PH06].

To analyse social judgments, Bless et al. [BFS04] introduced an information processing framework, which takes a stimulus event as input and results in a behavioural response after traversing the single steps of the sequence. To derive a final judgment, the process comprises the steps of *perception*, *categorisation*, *memory processing*, and *information integration*. Although all steps are essential to derive an appropriate decision, the emphasis of each step depends on the characteristics of the judged situation. Figure 2.2 visualises the steps of the information processing sequence based on a soccer referee's decision-making task. The following paragraphs describe each step involved in this process in more detail.

Perception: The information processing sequence starts with the perception of a stimulus, such as the observation of a tackling situation in soccer. To derive an accurate decision, all stimuli relevant to judge the performance need to be processed. Due to the reason that the human capacity to process information is limited, a judge needs to determine which information is taken into account in the subsequent information processing step. The perception of an event is strongly influenced by the referee's viewing position from which the performance of the athlete or team is observed [PH06]. By evaluating 200 offside situations, Oudejans et al. [OVB+00] analysed errors of soccer line referees' offside decisions. The results suggest that errors are often caused by the relative optical projection of the players on the referee's retina.

Categorisation : In the categorisation step, the judge interprets and encodes the perceived information by relating it to the prior knowledge stored in memory. Therefore,

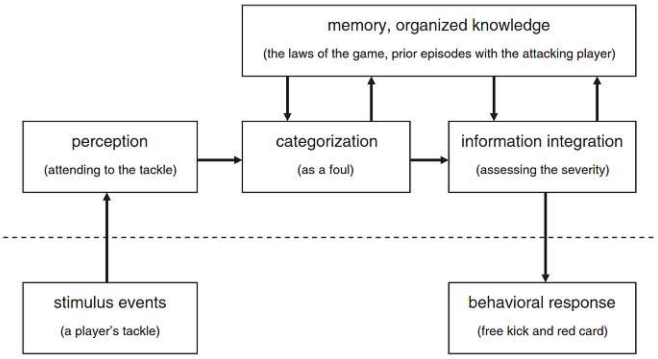


Figure 2.2: Social information processing sequence according to Bless et al. [BFS04] exemplified by judging a tackling situation in soccer by Plessner and Haar [PH06].

prior knowledge about the rules of the sport is a prerequisite to accurately judging the performance of athletes [PH06]. As a widely acknowledged assumption in social cognition, social knowledge is organised in complex structures interconnected by associative networks [BFS04]. According to Higgins [Hig96], knowledge activation is determined by its applicability and accessibility. While more chronologically accessible knowledge units are more likely to be used, knowledge units that were recently or frequently activated in the current context might be used instead. By analysing the effect of a team's uniform colour on referees' judgements, Frank und Gilovich [FG88] demonstrate that environmental cues irrelevant to objective judgements can activate knowledge. The results of this study show that teams in the National Football League (NFL) and National Hockey League (NHL) wearing black dresses were more frequently penalised than teams wearing other colours.

Information Integration: After the athlete's performance has been perceived and encoded, a final judgement is derived by incorporating information retrieved from the memory. Like in the step of perception, ideally, all relevant information is considered to make an accurate judgment. However, to overcome the complexity and the concomitant time pressure of the judgment situation, judges frequently use shortcuts which might not reflect the actual athlete's or team's performance [PH06]. An example in this area is the impact of crowd noise on the decision of potential foul situations in soccer. Neville et al. [NBMW02] examine this influencing factor by conducting a video-based experiment in which referees were assigned to one of two groups and prompted to assess potential foul situations. While one group observed the situations with crowd noise, the other group watched the video scenes without sound. The results showed that the presence of crowd noise led to more uncertainty in the referees' decisions and fewer fouls awarded against the home teams. Thus, crowd noise might contribute to the phenomenon of home advantage.

By taking a social cognitive perspective, the presented approach serves as a framework to understand the complex process of sports officials' decision-making and the variety of influences it is subjected to. Based on these foundations, measures to improve decision-

making skills can be developed by selectively addressing relevant steps in the information processing sequence specific to the demands of the judged situation [PH06]. While video-based feedback training can be applied to overcome deficiencies originating from retrieving improper knowledge structures in the categorisation step [SPKB11], specific training programs addressing the step of perception can be applied to compensate perceptual illusions [PWS+16].

2.1.4 Intuitive Decision-Making in Multi-Cue Environments

As outlined in the section above, the decision-making process comprises the steps of *perception*, *categorisation*, *memory processing*, and *information integration*. However, the emphasis of each step depends on the type of the judged situation [BFS04]. For example, Schweizer et al. [SPKB11] outline the importance of the categorisation step for judging foul/no-foul situations in soccer. Due to the similarities in judging contact situations in soccer and martial arts, the applied theoretical considerations might also apply to the training of referees in martial arts.

For the pragmatic modelling of the soccer referee training platform by Schweizer et al. [SPKB11], the basic ideas of social judgments were applied to the judgment of contact situations in soccer. By illustrating social judgments with Brunswik's Lens model [Bru52], the judgment of a situation is based on proximal and distal events. While proximal variables are directly observable, distal variables need to be inferred using cognitive strategies. In this regard, a contact situation in soccer is considered a set of proximal cues, where each cue is differently significant for the accuracy of the decision [BPS09].

As a learning paradigm to build the training platform for soccer referees mentioned before, the concept of *multiple-cue probability learning* was applied [BPS09]. In this learning framework, predictions and judgments are derived based on learning probabilistic relations between available information and an outcome. Learning in multiple-cue probability is based on repeated exposure to probabilistic information. Therein, the provided feedback about the correctness of judgments aims to build cue-outcome relations to improve the accuracy of judgments. Learning systems based on the integration of probabilistic information are called implicit learning. In contrast to explicit (or declarative) learning requiring awareness about the analytical processing, implicit (or procedural) learning does not require conscious control about the underlying reasoning [LNKS06].

A standard instrument to assess learning effectiveness in multiple-cue learning environments is the weather prediction task [KSG94], in which subjects learn to predict a binary outcome (sunshine or rain) by being repeatedly exposed to a combination of four different cards. The binary outcome is determined by the presented pattern, which consists of one to three cards, with each card independently contributing to the outcome with a fixed probability. Participants are supposed to gradually learn cue-outcome relations by completing a series of trials that include the following activities: (i) inspecting the presented pattern of cards, (ii) predicting the weather, and (iii) receiving immediate feedback on the prediction. Although not confirmed by Lagnado et al. [LNKS06], recent

research suggests that the performance of individuals in such learning environments does not necessarily reflect the individual's knowledge about the achievement of the task and the underlying decision rules [GSM02].

To derive decisions by integrating multiple cues under time pressure, recent research suggests that people use intuitive processing rather than deliberate processing. While intuitive processing is described as an automatic and effortless way to process information opaque to the decision-maker, deliberate processing is described as an effortful and reflective way to integrate information via a serial and rule-based approach by applying deductive reasoning. In addition, intuitive processing allows the parallel processing of information, which enables quick derivation of decisions [GB08, SPKB11].

According to Hogarth [Hog11], intuitions can be trained in learning environments where learners are provided with immediate, accurate, and relevant feedback. The learning environment must be representative for situations in which the trained intuitions will be applied. Similar to the approach of multiple-cue probability learning, the learning of intuitions does not require explicit feedback, explanations, or conscious awareness about the learning situation [SPKB11, Hog11].

2.2 Serious Games

Playing and learning are often perceived as opposites in public discourse. Although both activities comprise tasks which can be summarised as long, complex and challenging to master, they are contrarily perceived in terms of attractiveness. While resolving challenges in a game is an intrinsically motivated activity associated with freedom, joy and diversion, learning is associated with work, effort and concentration. One reason for the attractiveness of playing digital games over learning is their entertaining nature, created through interactivity that encourages a sense of self-efficacy. Using games as a medium to facilitate educational content can overcome the weaknesses associated with learning and turn it into a joyful and intrinsically motivating experience [BB10].

By combining game elements and learning in the right balance [BB10], serious games aim to achieve at least one other goal besides entertainment [DGEW16]. While no universal definitions of the term serious game exist, most authors agree that serious games are intended for purposes other than sole entertainment [SJB07].

Existing definitions in the literature require different properties that a serious game must meet. Due to the multitude of definitions and the resulting fuzziness, the spectrum of games being classifiable as serious games differs hugely, making it difficult to determine whether a game is serious. The serious aspect of a game can be determined by the perspectives of developers or players. By characterising a serious game by a player's intention, every digital game could become serious as long as it contributes to acquiring specific skills. An example is the use of an ego shooter game to improve reaction time [DGEW16]. A more general definition is provided by Abt [Abt70], which is not limited

to digital games, but covers games in general. By demanding a ranking between entertainment and learning, Michael and Chen [MC05] define the term serious game narrower by considering games, whose primary goal is education rather than entertainment.

In the subsequent sections, specific topics around serious games are discussed. Then, after discussing related concepts and introducing classification frameworks, design considerations for serious games are presented.

2.2.1 Serious Games and Related Concepts

As defined in the section above, serious games aim to improve the learning experience by blending entertainment and education. Apart from serious games, other media-based concepts also aim to facilitate knowledge and skills through similar approaches. Although concepts like edutainment, e-learning, and game-based learning share commonalities with serious games, they can be distinguished in certain aspects [RR09]. By describing the overlaps, distinctions and relations between these concepts, serious games are put in a broader context, contributing to a sharper categorisation of learning applications. Figure 2.3 visualises the relations between serious games and similar concepts in the form of a Venn diagram.

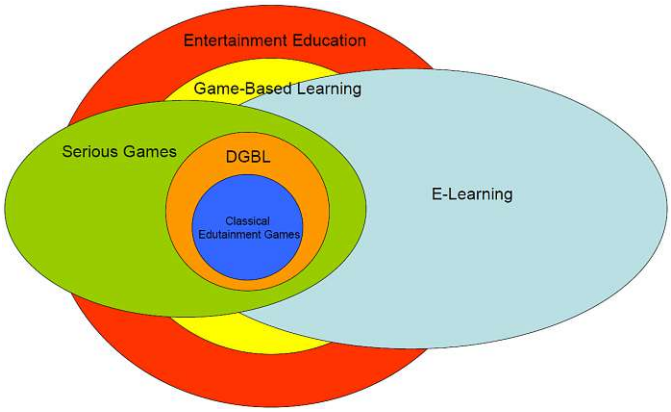


Figure 2.3: Serious games and their relation to similar concepts [BB10].

edutainment is a portmanteau word combining the terms education and entertainment. By not being limited to video games, it comprises any kind of education with entertaining elements [SJB07]. Edutainment applications are characterised by their age group, their teaching approach, and the way entertainment and education are blended. While serious games target all age groups, edutainment applications mainly focus on children in primary and secondary education. The teaching approach of edutainment applications focuses on teaching curricular textbook knowledge by additively providing entertaining elements for motivation. Although the terms edutainment and serious games are sometimes used as synonyms, most authors agree that edutainment applications are a subset of serious games [BB10]. However, due to the conveyance of learning content by conventional

learning approaches and the poor game experience, edutainment applications are heavily criticised [EN11, p. 18].

The model of *e-learning* comprises various approaches such as computer-enhanced learning, interactive technology, and distance learning in general [SJB07]. It comprises any form of computer-based learning that does not require enriching learning with entertaining elements. Due to their accessibility, flexibility and asynchronous nature, e-learning approaches enable learners to consume the provided learning content remotely. As some of these characteristics also apply to serious games, these categories are partly intersecting [BB10].

Game-based learning is a pedagogical approach aiming to reach certain learning outcomes by utilising games. It aims to create a joyful and motivating learning environment by combining game elements and learning [ALS18]. While serious games are applied in various domains, the approach of game-based learning is limited to the area of education [BB10]. *Digital game-based learning* is the subset of game-based learning including digital games only [SJB07].

2.2.2 Classifications of Serious Games

The literature proposes different approaches to classify serious games based on several dimensions. Thus, they contribute to understanding serious games and positioning them in a larger context. Based on the analysis of 612 serious games, Ratan and Ritterfeld [RR09] propose a classification system consisting of the four dimensions primary education content, primary learning principle, target age group and platform. The structured approach to synthesise categories based on collected descriptions of existing serious games allowed concluding the frequency distribution of the identified categories.

The following paragraph enumerates and explains the four dimensions of classifying serious games:

- **Primary education content:** This category is defined by the primary educational content, which the serious game aims to facilitate in addition to pure entertainment. It includes academic education, social change, occupation, health, military, and marketing. The vast majority of the analysed serious games can be assigned to the category of academic education (63 %).
- **Primary learning principle:** Serious games provide the opportunity to facilitate educational content through various methods such as exploration, experimentation, and problem-solving. This dimension is characterised by the kind of way how serious games attempt to convey skills, knowledge or ideas to the players. It includes practising skills, acquiring knowledge through exploration, cognitive problem-solving, and social problem-solving. Most of the analysed serious games had the primary learning principle of practising skills (48 %).

Table 2.1: Label-based classification system by Breuer and Bente [BB10].

Dimension	Exemplary Labels
Platform	Personal Computer, Nintendo Wii, Mobile Phone
Subject Matter	World War II, Sustainable development, Physics
Learning Goals	Language skills, historical facts, environmental awareness
Learning Principles	Rote memorisation, exploration, observational learning
Target audience	High school children, nurses, law students, pre-schoolers
Interaction mode(s)	Multiplayer, Co-Tutoring, single player
Application area	Academic education, private use, professional training
Controls/Interfaces	Gamepad controlled, mouse & keyboard, Wii balance board
Common gaming labels	Puzzle, action, role-play, simulation, card game, quiz

- **Target age group:** This category classifies serious games in target groups according to the age of the intended players. The classes consist of (i) preschool and below, (ii) elementary school, (iii) middle and high school, and (iv) the very general group of college, adults and seniors. The most frequently occurring categories within the analysed set of serious games were the classes (ii) and (iii), with a coverage of 39 %.
- **Platform:** This category describes the platform on which the serious game is played. It distinguishes whether serious games are played on a personal computer or another platform, such as a gaming console. The vast majority of the analysed serious games were executable on a personal computer (90 %).

The prevailing category in the dimension of primary education content is the field of academic education (63 %). Among the serious games classified as academic education, the dominant learning principle is skill practice.

Breuer and Bente [BB10] propose an alternative classification system based on the one of Ratan and Ritterfeld [RR09] and improves it in terms of completeness, flexibility and mutual exclusiveness. It does not only consider games explicitly designed as serious games but also COTS games (Commercial Off-the-Shelf Games) that can be used for a serious purpose. Rather than classifying a serious game by assigning a single category to each dimension, the system is based on flexible labels which can be assigned to nine dimensions. An initial set of exemplary labels is proposed for each dimension, which can be extended based on future developments. Table 2.1 shows the nine dimensions and a subset of pre-defined labels suggested by the authors.

Motivated by the significant growth of serious games, Laamari et al. [LEES14] introduced a taxonomy to classify serious games. The resulting classification framework consists of five dimensions activity, modality, interaction style, environment, and application area. The dimension of *activity* refers to the activity the player is requested to perform in the serious game. It covers the spectrum of physical, psychological or mental activities. The

serious game's *modality* describes how information is communicated from the digital game to the player. It describes how the player perceives the serious game considering visual, auditory, haptic and olfactory stimuli. Choosing proper modalities can enhance the learning experience and increase the players' motivation. The *interaction style* describes how the player interacts with the serious game. This includes traditional interfaces like keyboard, mouse or joystick as well as more-sophisticated interfaces such as brain interface, eye gaze, and movement tracing. Choosing suitable interaction styles in the game design ensures that the player adequately executes the requested tasks. Thus, it contributes to achieving the primary goal of the serious game. The *environment* of the serious game is described in different aspects. While the criteria 2D/3D, virtual reality, mixed reality and mobility describe the platform the serious game is executable, the criteria location awareness, online, and social presence refer to functional aspects. This dimension of *application area* refers to the domain in which the serious game is intended to be used, such as education, advertising, health care, well-being, cultural heritage, and interpersonal communication.

2.2.3 Flow, Game Flow and Dual Flow

Serious games have the potential to intrinsically motivate players and put them into a mental flow state where they are fully engaged by the game [DGEW16]. Csikszentmihalyi [Csi90] introduces a concept called flow, which describes a person's mental state of being fully immersed in an activity. Accordingly, he describes characteristics perceived by a person in the state of flow, which is described in the following enumeration:

1. **Challenging activities:** Being fully immersed in an activity requires a proper balance between task difficulty and the skills of the person performing the task, which is also referred to as the "golden ratio between challenges and skills". While a too-high task difficulty causes anxiety, performing a not challenging activity leads to boredom [Csi90, p. 49-53].

2. **Merging of Action and Awareness:** This characterises the situation where a person's attention is fully dedicated to the performed activity. By not being distracted by thoughts not relevant to the completion of the task, the execution of the activity becomes "spontaneous and almost automatic" [Csi90, p. 53-54].

3. **Clear goals and feedback:** Apart from creative activities such as painting a picture or composing a song, where the goal is not always clear at the beginning, the importance of having clear goals and getting immediate feedback is crucial to enter the flow state. Feedback allows to adjust the actions properly and verify if the primary goal is met. Furthermore, as indicated in the first point of this enumeration, the goal needs to have an appropriate level of difficulty to involve a person in an activity [Csi90, p. 54-58].

4. **Concentration on the task:** Concentration is often disturbed by preoccupations and anxieties, which prevents a complete focus on the task. To allow total concentration on an activity, it needs to be shielded from interfering thoughts and worries, leaving no room for irrelevant information distracting from the current task [Csi90, p. 58-59].

5. **Sense of control:** Being immersed in an activity typically requires experiencing a sense of control. People in a flow state are not worried about losing control, even if the potential consequences of failing might be serious. The potentially addictive activity of gambling constitutes an exception that does not require a sense of control to experience flow, as players have no impact on the random outcome of the game [Csi90, p. 58-62].

6. **The Loss of Self-Consciousness:** In many situations of everyday life, people need to reflect upon their self-image to react to potential threats appropriately. However, the flow state does not leave room for people to muse about the image of themselves, as they are entirely immersed in the activity [Csi90, p. 62-66].

7. **The Transformation of Time:** When experiencing flow, people often feel a discrepancy between the perceived and actual progression of time. People in this state usually report that time passes faster. The freedom not to worry about time increases the joy perceived during a completely involving activity [Csi90, p. 62-66].

As indicated in the first point, an equilibrium between task difficulty and a person's skill level is crucial for reaching the mental state of flow. Figure 2.4 visualises the proper balance between challenge and skill (A1, A4), which must be regularly adapted in order to keep persons in the flow channel and avoid moving towards the zone of boredom (A2) respectively anxiety (A3) [Csi90, p. 72-74].

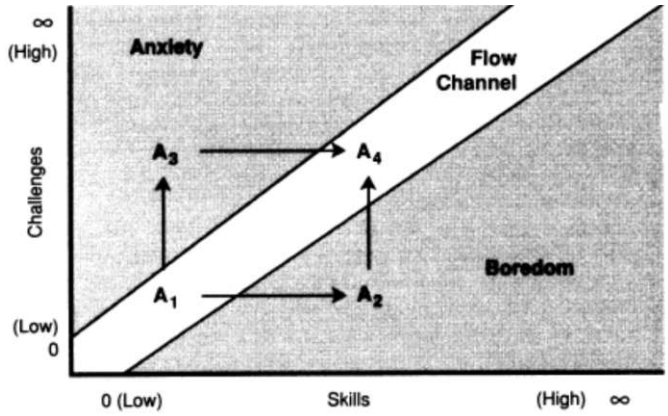


Figure 2.4: Flow channel depicting the proper balance between skill and difficulty [Csi90, p. 74].

The description of flow, according to Csikszentmihalyi [Csi90], is a rather general concept which can be triggered by a variety of activities such as games, art, and sports. By specialising the concept of flow to the domain of digital games, Sweetser and Wyeth [SW05] introduce the *game flow* model that includes the eight dimensions of concentration, challenge, player skills, control, clear goals, feedback, immersion, and social interaction. By transferring the concept of flow to the domain of serious games, *dual flow* describes the balance between task difficulty and player skills to fulfil the two main objectives of attractiveness and effectiveness [SHM09]. While initially described for the sub-category of exergames, Wiemeyer et al. [WNM6] generalise the concept of dual flow to the domain of serious games. Figure 2.5 illustrates the interplay between flow and efficiency.

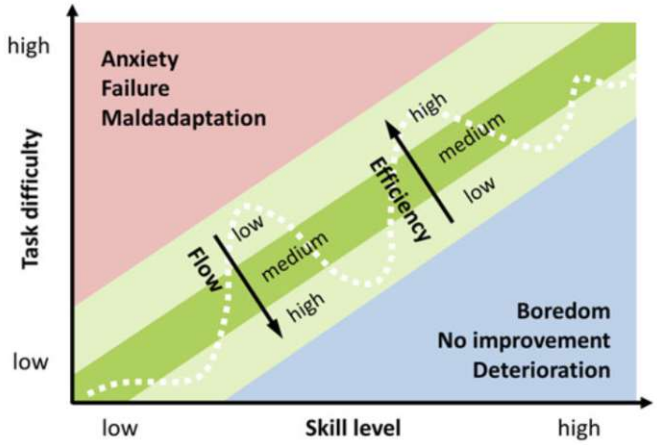


Figure 2.5: Dual flow as the balance between task difficulty and skill level [DGEW16, p. 11].

2.2.4 Adaption and Measurement in Serious Games

According to Göbel and Wendel [GW16], the term *adapting* refers to the process of changing something to fit new circumstances. Compared to personalisation, which refers to a one-time change of gaming aspects according to player preferences, adaptation involves a continuous game adjustment. In order to maintain the flow state, elements determining the game's difficulty need to be regularly adapted according to the progressing skills of the player.

The concept of player modelling describes capturing the player's characteristics as a decision basis to adapt the game appropriately. First, the initial version of the player model is built based on the user's preferences. Then, like the game, which is repeatedly adapted according to the player model, the model itself needs to be continuously updated to reflect the player's current state. In the domain of serious games, the concept of learner modelling is used to capture the performance of the player [GW16]. According to Kickmeier-Rust and Albert [MDKR12], indicators like score, completion rates, completion times, and success rates can be used to assess the skills and competencies of a learner.

2.3 Requirements Engineering

Requirements engineering is a systematic and disciplined approach to the specification and management of requirements, which aims to identify, document, and manage the stakeholders' expectations towards the target system to be built [PR15, p. 3-4]. This approach has been developed in order to address the lack of rigour frequently found in requirements documentation [PCY14, p. 66]. Requirements engineering has a significant impact on the success of customer-oriented projects. Deficiencies in this process result in unclear, incomplete, or wrong requirements, which can lead to the development of systems that do not fit customers' expectations. A common reason for incorrect requirements is the assumption by stakeholders that certain information is self-evident and can be communicated without being explicitly mentioned. This problem is caused by the heterogeneity of stakeholders in terms of experience and knowledge in specific domains. The importance of a high-qualitative requirements engineering process is further emphasised by the costs emerging from insufficient attention to this process. As the definition of complete requirements is the basis for system development, the effort required to fix poorly defined requirements increases as the software development project progresses. For example, fixing unclear requirements during development can be 20 times more expensive than fixing the same issue during requirements engineering. Therefore, an efficient requirements engineering process must be in place to detect deficiencies in early phases and prevent expensive change processes [PR15, p. 1-2].

2.3.1 Requirements Engineering in Software Process Models

Depending on the process model for developing the target system, requirements engineering is approached either as a self-contained phase or a continuous process. In classical process models, such as the Waterfall model, requirements engineering is performed as an initial phase in the project's life cycle. This approach attempts to ensure that all requirements are clearly defined before starting with design and development activities. In contrast, lightweight process models embed requirements engineering as a continuous process in all phases of system development [PR15, p. 4-5]. As classical process models freeze requirements specifications in early phases of the software development life cycle, such models are unsuitable if requirements are not well-understood or likely to change throughout the project [vC17].

2.3.2 Requirements Engineering Process

For capturing the stakeholders' expectations towards the target system, the requirements engineering process encompasses the elicitation, documentation, validation, and management of requirements. This process can be applied to define requirements on different levels of abstraction, such as stakeholder requirements, system requirements, and software requirements. The following paragraphs describe the activities involved in the requirements engineering process [PR15, p. 4].

Elicitation: The *elicitation* step focuses on gathering and refining requirements for the system to be developed from various sources by applying different techniques. This information can be collected from stakeholders, documents, and legacy systems, the primary requirements sources. The selection of appropriate elicitation techniques depends on various factors, such as the availability of resources and the requirements engineer's experience [PR15, p. 19-24]. When collecting requirements from stakeholders, it can be distinguished between explicit and tactic knowledge. While explicit knowledge is at the front of the stakeholders' minds and can be easily articulated, tactic knowledge refers to knowledge that stakeholders cannot easily articulate or explain. Thus, a challenge in requirements engineering is to make tactic knowledge more explicit [PCY14, p. 156-159].

Documentation: The *documentation* activity covers the description of elicited information using natural language or conceptual models, such as case diagrams, class diagrams, activity diagrams, or state diagrams. A requirement document must fulfil certain quality criteria regarding consistency, structure, extensibility, completeness, and traceability. Moreover, every single requirement must be (i) agreed with all relevant stakeholders, (ii) unambiguously documented, leaving no room for a different interpretation, (iii) comprehensible to each stakeholder, (iv) free of contradictions, (v) verifiable by performing certain measurements or tests on the implemented functionality, (vi) feasible with regard to organisational, legal, technical, and financial constraints, and (vii) traceable through origin, design, implementation, and test. [PR15, p. 33-47].

Validation: Documented requirements must be validated to ensure the quality criteria mentioned above are fulfilled. By reviewing requirements with stakeholders, the discrepancies between the requirements documented and the actual expectations can be identified and corrected accordingly in the early phases of requirements engineering. This results in approved requirements that can be used to start further development activities, such as design, implementation, and testing [PR15, p. 89-109].

Management: Requirements management encompasses activities ensuring the structure, traceability, maintainability, and evolution of requirements. Requirements documents and individual requirements must be managed throughout the entire software development life cycle. Requirements must be documented according to a well-defined set of attributes tailored to the needs of the individual project. Besides the requirement's name and description, a requirement is typically described by attributes such as unique identifier, author, sources, responsible persons, risk, and priority. Another area within requirements management is concerned with traceability throughout the system's life cycle. In general, three different classes of traceability relations can be distinguished: (i) traceability between requirements and their origin, (ii) traceability between requirements and artefacts resulting from subsequent development activities, and (iii) traceability between requirements. As requirements change over the system's life cycle, they need to be appropriately versioned to allow tracking of their evolution. Finally, a well-defined process must be in place for handling requirements changes based on justified decisions [PR15, p. 111-137].

2.3.3 Requirements Elicitation Techniques

According to Zowghi and Coulin [ZC05], requirements elicitation is acknowledged as an essential activity in the requirements engineering process. As a result, various techniques have been developed over the past decades to support requirements elicitation, many of which have emerged from social sciences. The following paragraphs describe some of the widely used techniques.

Interviews: Interviews are one of the most traditional techniques for eliciting requirements, which allows for efficiently gathering large amounts of data from individual stakeholders. Interviews can be divided into three large groups, which are characterised by the extent to which the interviewer adheres to a given structure: (i) unstructured interviews, (ii) structured interviews, and (iii) semi-structured interviews. By not following a predetermined list of questions, *unstructured interviews* enforce only limited control over the direction of the conversation. While unstructured interviews allow for exploring unfamiliar knowledge domains, they pose a risk of not being target-oriented by neglecting important issues while paying too much attention to unessential ones. In contrast, *structured interviews* are an effective and rigorous way to gather specific information from stakeholders based on a predefined set of questions. The success of such interviews depends on the interviewer's ability to formulate the right questions and ask them at the right time [ZC05]. As a mixture between unstructured and structured interviews, conversations in *semi-structured interviews* are guided based on the topics on the agenda rather than adhering to a strict set of predetermined questions. By combining open and closed questions, often accompanied by follow-up *why* and *how* questions, semi-structured interviews have the potential to reveal unforeseen issues from selected representatives of a target group [Ada15]. The data recorded from interviews with subject matter experts can be processed using qualitative content analysis, which aims to draw conclusions about certain aspects of communication by systematically analysing material emerging from any kind of communication [May15].

Questionnaires: Questionnaires provide an instrument to collect stakeholder information based on a predetermined set of open and closed questions. While they allow data collection from a broad range of stakeholders quickly, their usage does not allow for the generation of a deep level of knowledge. They also lack the opportunity to ask for clarifications, eliminate misunderstandings, or expand on new ideas. However, when the terms, concepts and boundaries of the domain are well established and the questions are focused, the conduction of questionnaires can be an effective method during the early stages of requirements elicitation based on which subsequent elicitation activities can be applied [ZC05]. More information about the design of questionnaires can be found in section 2.6.

Domain Analysis: By examining existing documentation and applications, domain analysis provides a useful way to elicit early requirements, gain domain knowledge, and identify reusable concepts or components. When the project scope involves replacing an existing system, studying design documents, instruction manuals, forms, and files

provides an essential source of requirements based on which complementary elicitation techniques such as observations and interviews can be applied. Besides projects involving replacing legacy systems, domain analysis can also be used to gather requirements in green field projects by inspecting similar or competitive applications [ZC05].

Prototyping: Prototypes are tangible artefacts supporting various stakeholders in envisioning the final system [BLM09]. This way, prototypes encourage stakeholders to participate in the development and refinement of requirements actively [ZC05]. Due to the importance of this method for collecting requirements and iteratively developing software systems, this method is described in a dedicated section (cf. section 2.4).

2.4 Prototyping

Prototypes are used in various industries to explore and limit product design uncertainties or identify problems in the production process before starting manufacturing on a large scale. Prototypes in software engineering are constructed as models, simulations or partial system implementations to demonstrate technical feasibility or as an instrument to determine user requirements [CV97]. In contrast to an abstract specification, prototypes are tangible artefacts supporting various stakeholders in envisioning the final system. By allowing the users to evaluate a representation of the final system in early design stages, prototypes provide a vital instrument in the approach of user-centred design [BLM09].

Prototypes can be used as learning vehicles facilitating a clear idea about the functionality to be expected in the target system. To unfold the full potential of prototypes for developers and users, they should address non-trivial and authentic problems relevant to the respective user group. They serve as a tool to converge to a viable software product by aligning the users' needs and the developers' possibilities. Involving users in early design phases and considering their feedback results in approved features that might increase the target system's acceptance. However, retrospectively changing already approved features without explicit user agreement can decrease the level of acceptance [Flo84].

While the term *prototype* refers to the designed artefact, *prototyping* describes the process of building them. The prototyping method attempts to mitigate deficiencies related to classical software development models, which assume that work steps can be specified before they are carried out. Software development models integrating the prototyping method can reduce the risk of incomplete specifications by dynamically reacting to changing user requirements. By promoting the decomposition of complex and ill-defined problems into several small parts, prototyping can help to build and refine applications that meet user or market expectations. Furthermore, the use of prototyping can help to (i) reduce costs, (ii) foster communication, (iii) analyse technical feasibility, (iv) improve risk management, and (v) involve users in the development process [CV97].

This section provides an overview of the method of prototyping in software engineering. After describing characterising dimensions of prototypes, the prototyping process and

construction techniques are outlined. The last section of this chapter examines different prototyping approaches with respect to the purpose they are supposed to fulfil.

2.4.1 Prototype Dimensions

Determined by the goal a prototype is supposed to fulfil, it needs to be constructed to meet specific quality criteria. While early design prototypes may suffice to show sketches of the prospective user interface, prototypes in later phases may require a representation and behaviour similar to the final system. Prototypes can be characterised along the four dimensions of representation, precision, interactivity, and evolution. Depending on the purpose the prototype is supposed to fulfil, the extent of these dimensions must be appropriately balanced [BLM09].

The dimension of *representation* refers to the medium of the prototype. The two basic forms of representation are offline prototypes and online prototypes. Offline prototypes do not require a computer and can be created quickly in early design stages without requiring skills in particular programming languages. Due to their low production costs, they enable rapid iteration cycles and the examination of multiple design alternatives. By not being limited by technical constraints, offline prototypes foster creativity and increase the number of generated ideas. Apart from these benefits, offline prototypes sometimes reach their limits when evaluating design ideas requiring immediate feedback to user inputs or dynamic visualisations. Online prototypes that can be executed on a computer are more appropriate in this case. Online prototypes are usually more expensive than paper prototypes, which decelerates the iterative design cycles. Rather than being used to examine various design ideas in early design stages, they are usually applied when the basic design strategy has already been determined [BLM09].

The *precision* of prototypes determines the level of required details the prototype needs to incorporate to fulfil its purpose. During the design process, the level of detail the prototype covers usually increases. Online prototypes generally have a higher level of precision than offline prototypes. To avoid facilitating the impression of high precision in early prototypes, designers often use offline prototype techniques such as paper sketches, whose representation is implicitly perceived as imprecise. The level of precision determines the balance between the elements included and (intentionally) omitted by the prototype. While the details covered by the prototype are subject to evaluation, the omitted elements are subject to discussion and exploration of the design space. The precision of a prototype must not be mixed up with its *fidelity* [BLM09]. While *precision* solely refers to the content of the prototype, *fidelity* additionally refers to the characteristics of representation, interactivity, and functionality [RSI96].

The dimension of *interactivity* determines the degree to which the prototype is able to respond to user inputs. The interactivity of a prototype is not limited by its representation or precision. Therefore, imprecise offline prototypes comprising a series of paper sketches can be constructed highly interactive by applying a person who plays the system's role,

presenting the respective cards in response to the user's request. Interactive online prototypes can be constructed so that only a subset of the functionality is implemented while a person simulates the missing parts. Independent of the techniques used, the user must experience the prototype as interactive. Depending on their level of interactivity, prototypes can be classified as fixed, fixed-path, or open prototypes. While *fixed prototypes* support no interactivity at all, *fixed-path prototypes* allow interactions along predefined scenarios. By allowing the user to explore certain functionalities interactively, *open prototypes* provide insights about how the user works with the system [BLM09].

The dimension of *evolution* refers to the intended life span of the prototype. The evolution of a prototype must be supported by applying appropriate construction techniques. With reference to the intended longevity of the prototype, prototypes can be distinguished between rapid, iterative, and evolutionary prototypes. *Rapid prototypes* can be produced at low costs and are characterised by their ability to explore various interaction styles in early design phases. After their evaluation, they are thrown away. *Iterative prototypes* are gradually refined in specific dimensions like precision and interactivity. A specialisation of iterative prototypes is depicted by the class of *evolutionary prototypes*, in which the prototype gradually develops into the target system [BLM09]. Closely related to the evolution and life span of a prototype, Floyd [Flo84] distinguishes the approaches of *exploratory prototyping*, *experimental prototyping*, and *evolutionary prototyping* [Flo84], which are further described in section 2.4.3. Among other characteristics, these prototyping classes also indicate the degree to which they are expected to be integrated into the target system.

2.4.2 Prototyping Process

The prototyping process can be described as a component of the software development methodology, applicable throughout all phases of the software development life cycle. Integrating prototyping in the software development process enables the production of early available artefacts demonstrating relevant parts of the target system and introduces an instrument fostering communication and feedback between involved stakeholders. Prototypes should be designed to allow for quickly adding and modifying features to enable fast adaption cycles in response to the collected feedback. The prototyping process includes the four steps of functional selection, construction, evaluation, and further use [Flo84].

Functional Selection

The first step *functional selection* determines the functional scope of the prototype and distinguishes two basic approaches: horizontal and vertical prototyping. Horizontal prototyping covers the implementation of specific individual layers, such as the user interface layer. Instead of implementing the functions in depth, a part of their effect is omitted, or simulated [Flo84]. By using appropriate tools for screen design, horizontal

prototypes can be produced fast, which allows the users to get an idea about the structure of the user interface in the early phases of the project [Nie93, p. 95].

On the other hand, vertical prototyping refers to the implementation of selected parts of the target system through all layers [BKKZ90]. Although vertical prototyping covers a limited set of features, it allows a thorough evaluation of certain work steps under realistic circumstances. A combination of horizontal and vertical prototyping is called *scenario*, in which a specific set of features is implemented with a reduced functionality allowing the prospective users to explore selected paths of the system [Nie93, p. 95]. Figure 2.6 visualises the comparison between the horizontal and vertical selection of features.

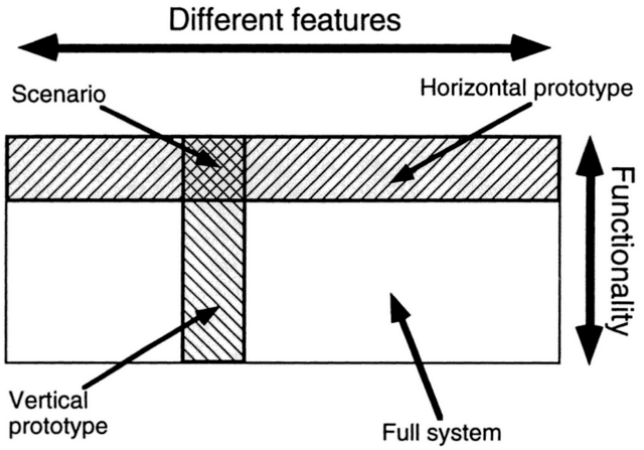


Figure 2.6: Feature scope of horizontal and vertical prototypes [Nie93, p. 94].

Construction

The step of *construction* covers the development of the actual prototype covering the previously selected set of features. By using appropriate tools and techniques to construct the prototype, the effort should be kept small. Since the purpose of the prototype is usually its evaluation rather than its long-term use, quality criteria such as reliability, security, or efficiency are negligible unless they are the subject of evaluation [Flo84].

Depending on the purpose of the prototype and its demands on the dimensions of representation, precision, interactivity, and evolution, different prototype construction techniques are applicable. By distinguishing prototypes with respect to the dimension of evolution, the following paragraphs outline construction techniques and principles applicable for rapid, iterative, and evolutionary prototypes [BLM09].

As *rapid prototypes* are inexpensive and easy to produce, they provide an instrument to quickly evaluate various interaction types in early design stages before they are thrown away. The degree of precision and interactivity impacts the time required to construct the prototype. Rapid prototypes can be created by offline and online techniques, whereas

online techniques allow for higher precision. Drawing sketches on paper, designing mock-ups tailored to specific physical devices, and producing videos containing a sequence of paper sketches are appropriate techniques for constructing offline prototypes. Figure 2.7 shows an example of a mock-up testing the user's interaction with the user interface on a handheld device. Interactivity can be simulated by the "Wizard-of-Oz" technique, where a designer plays the role of the system by manually presenting the respective content in response to user inputs. A more advanced approach for rapid prototyping is the usage of scripting languages to build interactive representations of the target system, which behave similarly to the target system [BLM09].

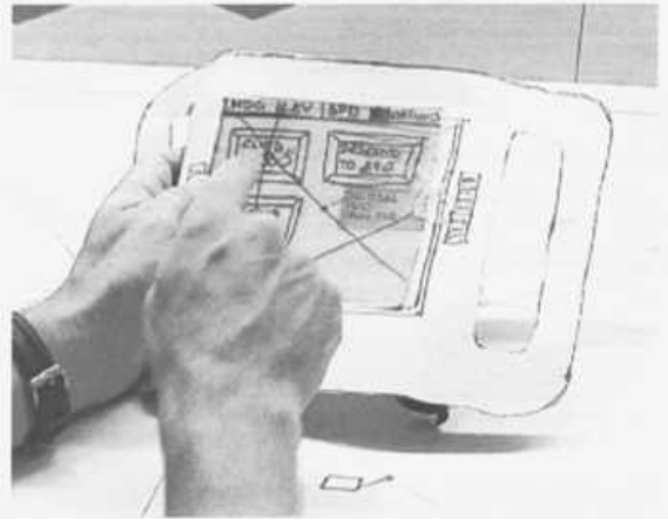
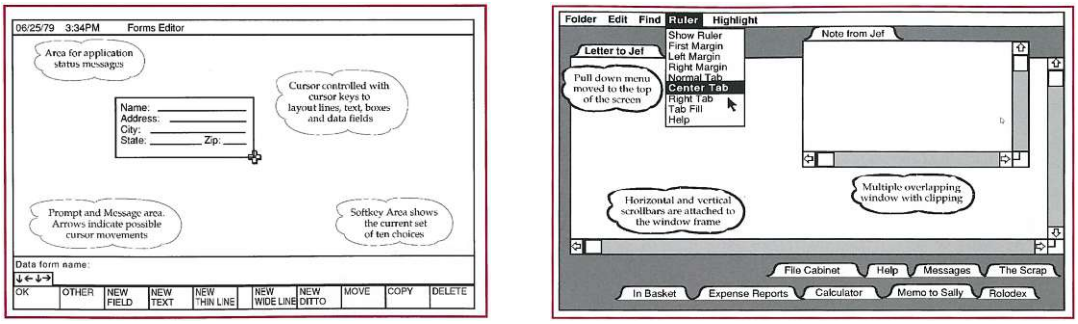


Figure 2.7: Mockup testing the user interface on a handheld device [BLM09, p. 10].

To allow for the creation of advanced *iterative* and *evolutionary* prototypes, tools supporting the composition of the graphical user interface can be used. Appropriate user interface toolkits support the development by providing standard widgets and the configuration of their behaviour in response to user inputs. Using dedicated user interface builders allows developers to interactively construct the user interface by arranging widgets on the screen. As *evolutionary prototypes* gradually evolve into the final system, they are additionally expected to fulfil non-functional qualities such as maintainability and performance. Therefore, they must be developed by following architectural principles supporting the evaluation of design alternatives at reasonable costs while still fulfilling the quality criteria mentioned above [BLM09]. Figure 2.8 illustrates the evolution of the Apple Lisa user interface from July 1979 to October 1980 [PKL97].

Evaluation & Further Use

After the construction of the prototype, the *evaluation* step retrieves feedback from relevant user groups, which is used as input for further development. The evaluation should be based on a protocol specifying the evaluation criteria and the sequence of



(a) Apple Lisa July 1979. (b) Apple Lisa October 1980.

Figure 2.8: Evolutionary prototypes of the Apple Lisa user interface [PKL97].

work steps to be performed. In addition, the planning of the evaluation also comprises organisational agendas to ensure the participants involved in the evaluation process are available and trained beforehand [Flo84].

By asking the users to perform certain scenarios under controlled conditions, this step of the prototyping process examines how the users experience different design alternatives. As in usability studies, the evaluation can include quantitative and qualitative methods [BLM09, SS17]. Commonly used methods are usability testing, think-aloud protocol, interviews, and questionnaires [SS17]. Often standardised questionnaires are applied, such as the *System Usability Scale* [Bro96] comprising ten questions to measure the usability of products in a technology-agnostic way [KB13]. In addition, metrics such as task completion rate, elapsed time, and the degree of required help can be recorded and used for benchmarking purposes [SS17].

The *further use* of the produced prototype is determined by the dimension of *evolution* [BLM09]. The reusability of the prototype *in* or *as* the final system depends on the gathered results in the evaluation step and the possibilities of the production environment where its functionality is supposed to be used. It either solely serves as a learning vehicle thrown away after the evaluation or is entirely or partially integrated into the target system [Flo84].

2.4.3 Prototyping Approaches

Depending on the goal the prototype aims to achieve, Floyd [Flo84] distinguishes the classes of *exploratory prototyping*, *experimental prototyping*, and *evolutionary prototyping*. Although these classes are associated with specific characteristics, the borderline between them is fuzzy and a clear distinction is not always possible. The following paragraphs provide an overview of these classes by describing their general purpose and characteristics with respect to (i) the relation to the phase-oriented software development model and (ii) the relation between the prototype and the target system.

Exploratory Prototyping

The purpose of *exploratory prototyping* is to mitigate communication problems between prospective users and developers. By filling the knowledge gap caused by the developers' potential lack of domain knowledge and the users' limitations in grasping the capabilities of the target system, a practical prototype demonstration can contribute to creative collaboration and the generation of good ideas [Flo84]. A possible approach in the class of exploratory prototyping is the method of *rapid prototyping* described earlier in this chapter [CV97].

This prototyping approach is used to clarify desirable features of the target system by discussing alternative solutions for given use cases. Thus, it aims to improve the early phases of the software development life cycle, including requirements engineering and functional analysis. Due to the goal of exploratory prototypes to gather ideas rather than being integrated into the target system, these prototypes are often poorly designed and thrown away after fulfilling their purpose [Flo84]. Thus, they have low demands on the dimension of evolution [BLM09].

An interdependent relationship exists between the users, the prototype, and the target system. On the one hand, the users' expectations towards the target system are influenced by the exposure to the prototype. On the other hand, the scope of the target system is influenced by the provided user feedback gathered with the help of the prototype [Flo84].

Experimental Prototyping

The focus of *experimental prototyping* is to evaluate the suitability of a proposed solution in an experimental setting before implementing it in the target system. Prototypes in this category can be evaluated with respect to many criteria, such as performance or feasibility. As this prototyping class is applicable in various areas, different approaches can be distinguished with respect to the scope of the implemented features. The method of *full functional simulation* implements all functionalities of the target system without attaching importance to efficiency, error handling or special cases. In the approach of *human interface simulation*, the user interface is implemented in its final form, while other system layers are mocked. By only implementing a few representative functions, *skeleton programming* allows the users to perform certain work steps in depth. The implemented features are supposed to serve as a blueprint, which allows users to imply the functionality of other work steps not covered by the prototype. The method of *partial function implementation* is used to examine specific parts of the system according to defined criteria, such as the number of consumed resources [Flo84].

After the initial requirements of the target system are specified, experimental prototyping can be applied in any phase of the classical phase-oriented software development model. Depending on the construction of the prototype, it is either thrown away or integrated into the target system. Due to financial constraints, experimental prototypes are prone to be adopted as production systems even though the quality criteria on the system are

not entirely fulfilled, which contradicts the actual intention of prototypes as a tool to increase the quality of the target system [Flo84].

Evolutionary Prototyping

Evolutionary prototyping is characterised by a long life span of a prototype and its ability to evolve into the target system [BLM09]. Instead of explicitly building a prototype as a learning vehicle to improve the quality of the target system, the emphasis of *evolutionary prototyping* is to gradually adapt the software product in response to changing requirements. Thus, it cannot be seen as prototyping in a narrower sense but more as "development in versions" [Flo84, p. 10]. While the approaches of exploratory and experimental prototyping primarily address communication problems and certain quality aspects, the evolutionary approach additionally aims to overcome the shortcoming of phase-oriented software development models, which assume that requirements are fixed at a particular stage in the life cycle. Evolutionary prototyping addresses this deficiency by developing the target system as a sequence of versions. Each version can be evaluated, and acts as a prototype for the successive version [Flo84].

The idea is to transform the linear ordering of development steps into consecutive development cycles. Depending on the activities performed in these cycles, the two basic approaches of incremental and evolutionary system development can be distinguished. *Incremental system development* refers to the gradual extension of the target system by primarily affecting the implementation phase. On the other hand, *evolutionary system development* considers software development as a sequence of successive cycles, whereas each cycle comprises the activities of design, implementation, and evaluation. This enables development in dynamic environments by flexibly responding to changing requirements instead of basing the implementation on a static specification captured at the beginning of the project. Despite the differences between incremental and evolutionary system development, both variants facilitate involving stakeholders in the software development process [Flo84].

Evolutionary prototyping can be adapted according to the situation's needs and combined with other prototyping approaches. For example, the development of a system can encompass cycles of exploratory and experimental prototyping during requirements analysis and design, followed by incremental system development to gradually develop the features of the target system [Flo84, BLM09].

2.5 Components of Web-based Systems & Eligible Technologies

From an architectural point of view, a web-based serious game like the one built in this study typically comprises components rendering the user interface, handling the business logic, and storing the data. The subsequent sections discuss possible frameworks and technologies to implement these components.

2.5.1 Web Application Frameworks

The goal of software frameworks is to eliminate repetitive operations by providing an abstraction layer containing solutions to common programming problems. A specialisation of software frameworks are web application frameworks, which encourage the development of web applications and services by providing a standard architecture that promotes consistency and predictability [VS11]. The purpose of web frameworks is to assist developers in building maintainable, configurable, and testable applications that provide services over the internet that can be consumed by multimodal clients [SB16].

A web application framework typically covers five tiers: client, web, business, data mapping, and data sources. The web tier handles incoming requests from the client and dispatches them to a controller that delegates the execution of the business logic to the business tier, which utilises a data mapper to access objects persisted in the data source. Once this operation is completed, the information is made available to the web tier, where a template engine generates the response to be returned to the client. Depending on the application's architecture, the response of the server either contains an HTML document or information wrapped into a structured data format such as JSON (JavaScript Object Notation) or XML (Extensible Markup Language) [SB16].

According to the Stackoverflow Developer Survey 2022 [staa], among the most popular web application frameworks are Node.js [nod], Django [djab], and Ruby on Rails [raia]. Figure 2.9 visualises the popularity of these frameworks based on the number of questions asked by the Stack Overflow community over the past few years [stab]. The trend shows that Node.js and Django have an active community, which has increased recently, while activity around Ruby on Rails has decreased. The subsequent sections outline these web application frameworks by describing their purpose, architecture, tiers and characteristic features.

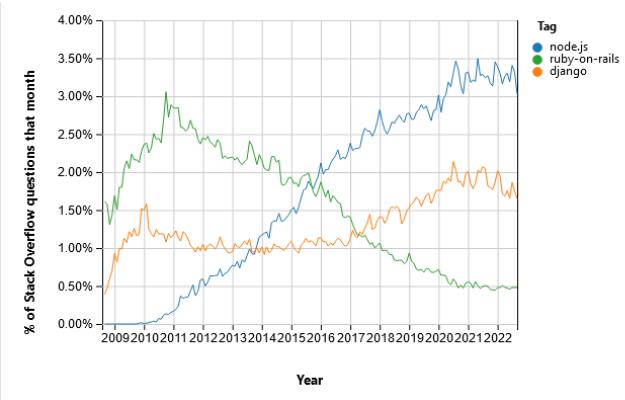


Figure 2.9: Stack Overflow trend of Node.js, Django, and Ruby on Rails [stab].

Node.js and Express

Express is a web application framework based on the open-source JavaScript run-time environment Node.js [nod]. It provides advanced HTTP routing and middleware utilities for creating robust API endpoints for web and mobile applications by using JavaScript [expa]. Incoming HTTP requests initiated by the client are matched and dispatched to the respective controller performing the business logic, which communicates with a database to perform the requested function. Once the business operation is performed, a template engine generates a valid HTML file by replacing the defined variables with actual values, which is returned to the client [SB16]. Express.js does not come with a template engine and an object-relational mapper (ORM) out-of-the-shelf. However, the Node.js ecosystem provides various template engines [expd] and ORM libraries [pri, seq] that are compatible with Express.js. A template engine can be skipped if Express.js is only used to expose API endpoints consumed by a single-page web application or mobile app that dynamically updates the user interface on the client-side [SB16].

Django

The Django framework [djab] supports developers in writing web applications in the programming language Python. While it follows the MVC (Model View Controller) architectural pattern, it describes itself as MTV (Model Template View) framework, in which the view corresponds to the controller and the template corresponds to the view according to classical naming conventions [djac]. Requests triggered by the client are handled by a Python module, maintaining the mapping between URL patterns and callback functions, so-called views. It matches the requested URL based on regular expressions and delegates the execution to the associated view, which acts as a controller to trigger the execution of the business logic. The execution of the business logic might require retrieving objects persisted in the database, which is supported by a built-in ORM that allows defining models and communicating with the database using Python code [djaa]. Once the business operation is completed, the template engine converts the template written in *Django template language* [djad] to an HTML file, which is returned to the client. As for Express.js, no templates are required in case Django is only used to expose API endpoints to client applications. In this case, the web tier returns the requested data in structured formats such as JSON or XML [SB16]. Apart from providing a framework for handling the request-response life cycle of web applications, Django allows to automatically create an administrative user interface based on the defined model. This allows users to comfortably create, update, and delete content in the database as soon as the models are defined [djaa].

Ruby on Rails

Ruby on Rails [raia] is a web application framework for the programming language Ruby, which aims to simplify the development of web applications based on the principles *Convention over Configuration* (CoC) and *Don't Repeat Yourself* (DRY). The CoC

principle refers to the adherence to conventions reflecting convenient best practices rather than requiring developers to configure every single aspect of the application. The DRY principle aims to increase maintainability and extensibility by avoiding writing information repeatedly. Like other frameworks, the separation of concerns is based on the MVC architectural pattern. Routes map incoming HTTP requests from the client to actions defined in the controller. The controller prepares the requested data and hands over the control flow to the respective view that uses the ERB (Embedded RuBy) template engine for generating the web page to be returned to the client [raib]. The communication with the database is supported by models following the Active Record pattern [Fow, act]. The framework provides scripts for generating models, views and controllers separately or all at once. An example is the *generate scaffold* script, which automatically creates a model, database migration, controller, routes, and views for a given entity by providing its attributes [raic]. This way, a simple web application allowing the user to navigate through the user interface to create, read, update, and delete objects can be automatically generated using a single command. For controllers created by a generator script, Rails automatically creates a JSON API that can be consumed by clients to retrieve data and render it to the user. Thus, the framework is not only applicable for developing web applications but also for developing endpoints consumable by any client application [SB16].

2.5.2 Client-side Web Frameworks

While the web application frameworks mentioned above work with a template engine supporting the creation of requested HTML pages, they can also be combined with client-side JavaScript frameworks to create highly interactive user interfaces [SB16]. Single-page web applications aim to bring the look and feel of native applications to the web browser by utilising JavaScript, HTML, and Cascading Style Sheets [Sco15, p. 3]. Compared to classic multi-page web applications reloading the entire user interface upon each request triggered by the user, single-page web applications asynchronously update individual elements of the user interface without refreshing the entire page. This way, single-page web applications increase the degree of interactivity, responsiveness, and user satisfaction [MvD07]. Dynamic web pages are generated on the client side based on asynchronously requesting data from the APIs built with the help of the web application frameworks mentioned above [SB16].

According to the Stackoverflow Developer Survey 2022 [staa] conducted in the period from May 11th, 2022 to June 1st, 2022, the most popular client-side frameworks are React.js [reaa], Angular [angc], and Vue [vuec]. These frameworks enable the construction of interactive user interfaces and are typically used to create single-page web applications. Figure 2.10 visualises the popularity of these frameworks based on the number of questions asked by the Stack Overflow community over the past few years [stab]. The trend shows that React.js surpassed the popularity of Angular in 2019 when the popularity of Angular started to decrease. The subsequent sections outline these three frameworks by describing their main characteristics.

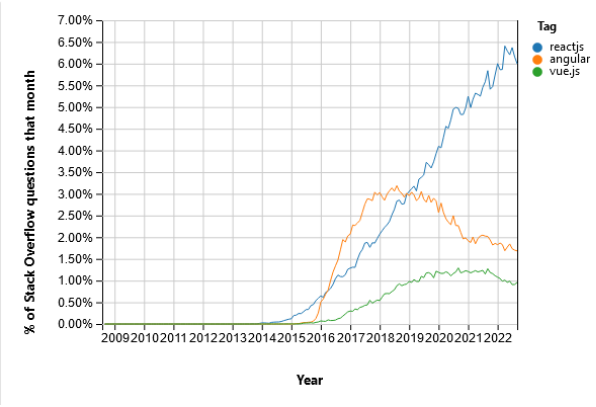


Figure 2.10: Stack Overflow trend of React.js, Angular, and Vue.js [stab].

React.js

React is a component-based JavaScript library that enables the creation of advanced user interfaces by composing multiple independent and reusable components managing their own state [reaa]. React is considered a lightweight framework that provides a basic set of features that can be selectively extended by third-party libraries developed by the React community [Vya22]. A React component can be defined as a JavaScript function (i.e. function components) or a JavaScript class (i.e. class components). In both variants, a component defines the HTML template and JavaScript functions determining its behaviour [reac]. With the introduction the JSX (JavaScript XML) as a syntax extension of JavaScript, a component's HTML template can be written in JavaScript and added to the DOM without the need to explicitly use the methods *createElement* and *appendChild* [JSX]. React uses the approach of one-way data binding: While changes in the component's state automatically trigger the update of its view, the data from the view needs to be actively read by attaching an event listener to the respective element in the document object model [Vya22].

React Native [read] is a framework based on React for developing mobile applications by using JavaScript. It allows the writing of native mobile applications for Android and iOS, where a large portion of the code can be reused between platforms. Like React, the user interfaces in React Native are composed of components written in JSX syntax, which allows embedding markup language in the JavaScript code. Compared to React, which uses HTML and CSS as a markup language for rendering views, React Native uses dedicated components, which are translated to native UI elements of the host platform [Eis15].

Angular

Angular [angc] is a component-based framework for building interactive and scalable web applications in TypeScript according to a Model-View-Controller architecture [Vya22].

It provides a set of built-in libraries comprising a wide range of features such as routing, forms management, and client-server communication. If required, the functional range can be extended by numerous first-party and third-party libraries [angd]. Angular web applications are composed of reusable components that include HTML templates, styles, and Typescript classes defining the component's behaviour. This way, the framework ensures the separation of concerns between the application's logic and its presentation [anga]. By extending HTML with a special syntax, Angular allows to insert dynamic values from the component [angb]. Angular supports a two-way data binding, where changes in the model are automatically reflected in the view and vice versa, which negatively affects the performance [Vya22].

Vue.js

Vue [vuec] is a JavaScript framework that enables the development of user interfaces based on a declarative and component-based approach. In terms of functionality, Vue is somewhere between React and Angular. Since the number of packages available for Vue is less than that for React, choosing suitable packages for a given use case is more difficult [Vya22]. Vue is based on the two core principles of declarative rendering and reactivity. Declarative rendering refers to the extension of HTML with the help of a template syntax describing the HTML output based on the component's state. Reactivity refers to the automatic tracking and efficient update of the DOM as a response to changes in the component's state [vueb]. Like Angular, Vue also supports two-way data binding. This way, changes in the view are automatically reflected in the component's state and vice versa [Vya22]. Components in Vue are written in the form of so-called Single-File Components that wrap the component's logic, template, and styles in a single file [vueb]. Applications are structured in a tree of reusable components that can be arbitrarily nested [vuea].

2.5.3 Database Systems

Depending on the requirements the software application is supposed to fulfil, the selection of proper technology for storing data is essential. A key decision is whether to use a relational or NoSQL (Not Only SQL) database, which differs in various properties such as scalability, performance, and flexibility. Relational databases are designed to support vertical scaling, which requires an increase in resources on the server hosting the database in response to a rapidly growing amount of data. NoSQL databases, on the other hand, allow for horizontal scalability, where new nodes are added to increase data storage and processing power. Both database types follow different strategies to ensure certain properties of the stored data. While the priority of relational databases is to meet the ACID (Atomicity, Consistency, Isolation, Durability) properties that ensure data integrity and reliability, NoSQL databases follow the BASE (Basically Available, Soft state, Eventually consistent) principles that allow for a higher level of flexibility and support horizontal scalability. Another distinguishing feature of the two types of databases is the structural flexibility in organising the data. Data in relational databases

is structured according to a predefined schema. In contrast, NoSQL databases' dynamic schemas allow for storing structured, semi-structured, and unstructured data. In general, relational databases are recommended in case data integrity and consistency is important, while NoSQL databases provide better performance and scalability for rapidly growing data volumes. [SAZ+18].

According to the Stackoverflow Developer Survey 2022 [staa] conducted in the period from May 11th, 2022 to June 1st, 2022, the four most popular databases are the relational databases MySQL [mys], PostgreSQL [pos], and SQLite [sql] as well as the document-based database MongoDB [monb]. Data in MongoDB is structured in collections and documents, where each collection can contain multiple documents persisted in the form of JSON-like objects [monc]. While many document-based databases only provide a basic query language [SAZ+18], MongoDB comes with the sophisticated query language MQL (MongoDB Query Language) [mona], which allows to efficiently query data by using functions for filtering, joining, and aggregating the data set.

Figure 2.10 visualises the popularity of these databases based on the number of questions asked by the Stack Overflow community over the past few years [stab]. The trend shows that PostgreSQL and MongoDB have an active community, which has increased in recent years. On the other hand, although the community around MySQL still shows the highest activity of the four mentioned databases, it strongly decreased in the last few years. The activity around SQLite has only slightly decreased over the past years.

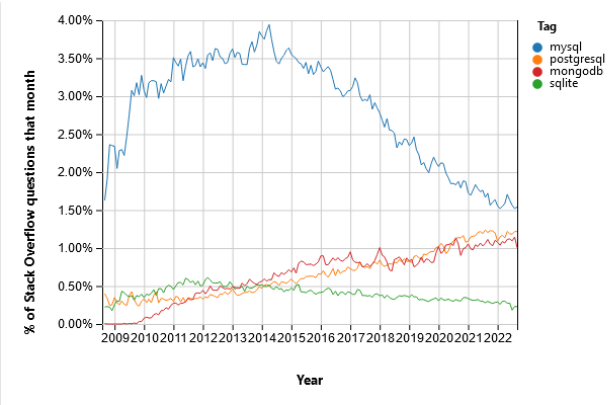


Figure 2.11: Stack Overflow trend of MySQL, PostgreSQL, SQLite, and MongoDB [stab].

2.6 Questionnaire Design

According to Sheatsley [She83, p. 200-201], questionnaire design is neither a science nor a technology guaranteeing the outcome of a high-quality questionnaire. However, it provides guidelines and best practices to avoid common pitfalls causing poor questionnaires. A good questionnaire is characterised by its ability to (i) meet the research objective, (ii) gather complete and accurate information and (iii) adhere to cost and time constraints.

To compose questionnaires meeting the research objective, they need to be designed by not steering results in a particular direction. Potential bias of researchers can lead to favouring less important variables while neglecting important ones and thus falsifying the outcome. To gather complete and accurate information from the respondents, the questionnaire needs to be structured and phrased to avoid misunderstandings, rejecting the question's premise or refusing the answer. Besides design-related qualities, the execution of the questionnaire needs to adhere to given cost and time constraints to ensure the research's feasibility and the timeliness of the gathered data [She83, p. 201-202].

To produce a questionnaire fulfilling these qualities, Sheatsley [She83, p. 202-203] proposes an iterative approach of (i) content definition, (ii) question drafting, (iii) question structuring, and (iv) pretesting. Thus, the first step in designing a questionnaire is clearly defining the required information needed to meet the survey's objectives, which contributes to selecting appropriate questions. However, achieving the survey's objectives depends not only on the selected questions but also on how they are structured, phrased, and tested. This section discusses aspects of questionnaire design that fulfil the quality criteria mentioned above. Based on scientific foundations, essential considerations and recommendations for the questionnaire design processes are described.

2.6.1 Questions Types

According to Krosnick et al. [KP09], questions in a questionnaire can be categorised into open and closed questions. Compared to open questions, where respondents are encouraged to provide answers in their own words, closed questions suggest a set of defined response options to be chosen by the respondent. Therefore, when deciding whether to use the variant of open or closed questions, their accompanying advantages and disadvantages must be considered with respect to their ability to achieve the questionnaire's objective.

While the benefit of open questions comes with their character of allowing respondents to express answers in their own words without being influenced by suggested response items, they depend on the respondent's articulation ability and require the researcher to interpret and cluster the responses correctly. In contrast, closed questions simplify the analysis process, as coded answers can be evaluated without needing to interpret and cluster the responses manually. However, closed questions restrict respondents from expressing themselves by a limited set of response categories, which might not reflect the exact shade of their answer [She83, p. 206-208]. The higher costs associated with evaluating open questions contribute to the broad usage of closed questions [KP09].

Aside from the strict distinction between open and closed questions, a combination of both variants is appropriate. By extending closed questions with an additional option to allow an open response, respondents can express themselves in case the preferred answer is unavailable in the suggested list of response items. However, the use of this approach must be carefully elaborated due to the tendency of people to select suggested answers and the disadvantages in terms of evaluation mentioned above [She83].

2.6.2 Phrasing Questions

The way questions are phrased contributes to the accuracy and validity of the produced data. Good questions must be (i) focused on the research objectives, (ii) phrased according to the respondent's level of knowledge and (iii) suitable for the administration method [She83, p. 205-206]. The following paragraph enumerates principles of question writing intending to avoid respondent confusion, misunderstanding, lack of comprehension and response bias [She83, p. 212-217].

Keep it simple: While questions are usually phrased by educated people familiar with specific terms and concepts in the affected domain, respondents might have a different educational background. Avoiding long words, technical terms, and complex phrases allows a broad audience to understand and properly answer the questions. A lack of understanding might lead to so-called "don't know" responses or the selection of an arbitrary answer instead of admitting that the question has not been understood.

Keep it short: To support the simplicity of questions, they should be kept as short as possible by not exceeding a maximum range of 25 words if possible.

Provide alternatives: When asking for the respondent's reaction, the wording of the question should always specify alternatives. It must be avoided to influence respondents by presenting a biased version of the question, as people tend to acknowledge with the phrased proposition. A good question conveys that any response is acceptable.

Loaded questions: Loaded questions attempt to elicit information by making participants agree with the question's assumption. Although there are specific cases for which loaded questions might be a viable strategy to elicit information, they should be avoided in general.

Double-barreled questions: One common pitfall is the use of double-barreled questions, where multiple positions are packed into a single question. Such questions cannot be correctly answered unless respondents agree or disagree with all positions in the same intensity. This problem can be easily solved by splitting the question to elicit the required information separately.

False premise: The false premise error relates to questions that are based on a premise the respondent might or might not agree with. If the respondent disagrees with the premise, the question cannot be answered correctly.

Intentions to act: Questions asking for the respondents' future behaviour are not expressive in general, as actions are influenced by many situational variables. A more promising approach is to derive the respondents' future behaviour based on questions querying their past and current attitudes.

Vague, ambiguous words: If terms used in questions are ambiguous, they can be interpreted in different ways. Therefore, items used in questions need to be precisely specified to query accurate information from the respondents.

Double negatives: A possible source of confusion is the formulation of questions asking for agreement or disagreement of a negated term.

2.6.3 Response Items Scales

The decision to use the format of closed questions requires the definition of appropriate response categories. The following paragraphs discuss various scale types as well as their applicability. Due to their importance for measuring respondents' attitudes, Likert scales are separately discussed in the subsequent section.

Nominal scales: Nominal scales allow classifying the sample into mutually exclusive categories, whereas no hierarchical relationship exists between the response items. Furthermore, no numerical differences exist between the single response categories. Examples of nominal variables are gender, religion or occupation [FM09].

Ordinal scales: Ordinal scales classify participants into hierarchically ordered categories. This scale type is applicable for not directly measurable variables like satisfaction, pain, or anxiety [FM09]. Response items on ordinal scales are ordered in terms of the < or <= relation, making them comparable. In contrast to interval scales, which are characterised by the property of equidistant item values, the difference between the values on ordinal scales is negligible [GMR07]. A special form of an ordinal scale that measures respondents' attitudes are Likert scales [Lik32], which are discussed in section 2.6.4.

Continuous Scales: Continuous scales can be further classified into interval and ratio scales and are measured by numeric values (e.g. weight in grams). In contrast to ordinal scales, the increments on continuous scales are equidistant. The difference between ratio and interval scales lies in the existence of an absolute zero point. While data in ratio scales is characterised by the existence of an absolute zero, this property does not exist for interval scales. Examples are length, height and weight for ratio scales, respectively temperature for interval scales [FM09].

2.6.4 Measuring Attitudes with Likert Scales

Many questions form natural dichotomies, where simple binary answers like "Yes or No" or "Approve or Disapprove" are sufficient to classify respondents into two groups. However, phrasing questions as natural dichotomies is not always the best choice, as they provide a poor response distribution. To address this problem, questions can be rephrased to allow answers on a scale covering the spectrum from total agreement to total disagreement. For example, a five-point scale allows a binary grouping into agreement and disagreement by aggregating the two agreement and disagreement positions. However, in addition, a finer-grained evaluation can be conducted if required [She83, p. 208-209].

The five-point scale mentioned in the paragraph above was developed by Renis Likert in *A technique for the measurement of attitudes* [Lik32] and is therefore commonly referred to as Likert scale. Therein, statements are presented to respondents, who are asked to choose among several grades on the scale representing their best-fitting attitude. Hence, these scales are often used in surveys for measuring the attitudes of respondents [GMR07]. The following paragraphs discuss important considerations for designing and interpreting Likert scales.

Labels and number of response items: Depending on the research's objective and population (general population versus student population), Weijters et al. [WCS10] propose a decision framework for the design of Likert scales concerning the labelling and number of response items. The framework suggests using 5- or 7-point scales for student populations and a 5-point scale for studies involving participants of the general population. The labelling of the points on the scale depends on the objective of the research. While a fully-labelled scale is suggested for newly developed scales and measurements of opinions, studies aiming to estimate linear relations are recommended to label the endpoints only.

Inclusion of Midpoint in Likert Scales: The use of a 5- or 7-point scale implies the existence of a neutral/midpoint element, which has been the subject of discussion for decades [WCS10]. Chyung et al. [CRSH17] state that adding a midpoint is no general decision but depends on the individual use case. They propose that a midpoint should be included in case (i) respondents are familiar with the surveyed topic and (ii) the scale is statistically analysed as an interval scale.

Interpretation of Likert Scales: As discussed by Göb et al. [GMR07], there are controversial points of view concerning the methods applicable for the statistical analysis of Likert scales. Concretely, the controversy lies between interpreting the Likert scale as an ordinal scale versus an interval scale. While it is stated that measuring respondents' attitudes requires assuming an ordinal scale interpretation, studies often apply statistical methods that are only applicable to continuous scales. This issue is also discussed by Svensson [Sve01], who states that one must be careful when applying statistical evaluation methods for qualitative variables measured by ordinal scales. Although the items on ordinal scales are often annotated with numerical labels, they only indicate an ordered structure, which limits the applicability of mathematical operations like building sums, differences, means or standard deviations. Even though interval interpretations of Likert scales are frequently applied in practice [PZB88, DELG85, WCT88], measuring attitudes suggests an ordinal scale interpretation [GMR07].

2.6.5 Questionnaire Structure

A questionnaire does not immediately start with a series of questions but first on-boards the participants. This is done by providing basic information about the survey's general purpose and content in a compact format. To eliminate possible reservations, a good introduction must at least explain (i) for whom the survey is conducted and (ii) the

purpose of the survey. This information is usually written on the questionnaire but can also be announced in advance. Another possibility is to place the introduction in the invitation to the survey [She83, p. 219].

After the respondents are informed about the purpose of the survey, the actual questions are stated. When composing a questionnaire, the structure and order of asking the questions are of special meaning. A general rule is to keep the opening question simple and non-threatening, as it sets the tone for the rest of the questionnaire and respondents are encouraged to continue. Otherwise, respondents might become suspicious, which increases the risk of a break-off or evasive responses [She83, p. 220].

The flow of questions should be arranged in a psychological order, where one question naturally leads to the next. Questions should be coherent with respect to the subject they are addressing, and hopping between topics should be avoided. If the questionnaire covers several topics, it is beneficial to announce new sections with a separate introduction. Within the respective topic, the flow of questions should start with general questions followed by more specific ones [She83, p. 221].

2.6.6 Pretesting

Pretesting the questionnaire aims to reveal major weaknesses and thus increases the quality of the questionnaire. The magnitude of pretests can range from testing questions in small groups up to extensive pilot studies aiming to test how the items behave statistically [She83, p. 225-227].

In [GM20], Geisen and Murphy review both classical and enhanced pretesting methods that emerged from the increased conduction of web and mobile surveys. Among other approaches, the classical pretesting method of *expert review* is described, in which the questionnaire is inspected by survey methodologists or subject matter experts. This pretesting method intends to assist the questionnaire writer in detecting common flaws like double-barreled questions and overlapping response options.

To increase the formality of the review process, it is suggested to follow a standardised protocol like the *question appraisal system* (QAS-99) introduced by Willis and Lessler [WL99]. This protocol aims to improve the efficiency of the questionnaire review process by providing a standardised protocol to detect problems in drafted questions and revise them accordingly. It provides a structured guideline focused on characteristics of questions that are likely to cause response errors. During the review process, each item in the questionnaire is systematically evaluated by the reviewer according to the steps defined in the guideline. The guideline consists of seven steps covering reading, instructions, clarity, assumptions, knowledge/memory, sensitivity/bias and response categories. Each topic is further subdivided into two to seven so-called problem types, which provide the basis for the evaluation and can be answered by either *yes* or *no*. In case *yes* is selected, an additional comment describing the reason for the problem must be added. Following this standard procedure, questions likely to cause problems can be detected and

improved accordingly. Dean et al. [DCM+07] introduce an extended version of QAS-99, the Question Appraisal System (QAS-04), which additionally considers problems arising in cross-cultural and cross-lingual contexts.

2.7 Descriptive and Inferential Statistical Methods

Both descriptive and inferential statistical methods can be applied to analyse data collected through surveys or field experiments. While descriptive analysis summarises the characteristics of the sample, inferential analysis allows generalising the observations by drawing further conclusions about a larger population [FM09].

2.7.1 Descriptive Methods

After collecting data from a sample, the first step is to summarise and describe important characteristics of the data [Sta20]. Gathered data can be meaningfully described by calculating a single index value representing the entire data. The central tendency is a measure describing the data set's centre point by a single representative value. Commonly used measures to indicate central tendency are mean, median and mode. The mean provides a representative average by considering every value in the observed data [Man11b]. In contrast, the median marks the middle position in the sorted data and thus is equivalent to the 50th percentile. The mode indicates the most frequently occurring value in the data set [Man11a].

The methods of frequency distribution, range, interquartile range, and standard deviation can be used to describe the degree of dispersion of a variable. While the frequency distribution indicates the number of occurrences per category, the range and the interquartile range show the distance between the minimum and maximum or the first and third quartile. The standard deviation indicates the average distance of each observed value to the mean of the entire dataset [FM09].

The applicability of descriptive statistical methods depends on the level of measurement of the given variable. In general, four levels of measurement can be distinguished: nominal, ordinal, interval and ratio [FM09]. Table 2.2 shows the applicability of descriptive methods for central tendency and dispersion dependent by scale type.

2.7.2 Inferential Methods

A statistical study typically focuses on a defined collection of objects that form a population of interest. In case the study obtains data for all objects in this population, it is called a census. Due to limited resources, a study usually collects data from a representative subset of the population, a sample. However, inferential statistical methods can be used to generalise the information obtained from the sample and draw conclusions about the entire population of interest [Sta20]. For example, inferential statistics can be used to examine whether differences between groups are solely due to

Table 2.2: Applicability of descriptive statistical methods by level of measurement [FM09].

Scale type	Central tendency	Dispersion
Nominal	Mode	Frequency distribution
Ordinal	Mode Median	Frequency distribution Range Interquartile range
Interval, Ratio	Mode Median Mean	Frequency distribution Range Interquartile range Standard deviation

the selected sample or whether they reflect actual differences between the populations represented by the groups [AT09].

Inferential statistics can be divided into parametric and non-parametric methods. By making certain assumptions about the measurement level, distribution, and other data properties, parametric tests are more powerful and sensitive than their non-parametric counterparts. A basic requirement of parametric tests is that variables must be measured at an interval or ratio level. Moreover, when parametric methods are used, it is assumed that the values of the examined variables are normally distributed in the population, subjects are randomly selected, and variations in results from each condition are homogeneous [BSW14]. Examples of parametric methods applicable to normally distributed data measured on interval or ratio scales are Student's t-test and ANOVA (analysis of variances) for testing differences between groups and Pearson's r for testing relationships between two variables [AT09].

If the requirements for the use of parametric tests are not met, equivalent non-parametric statistics can be used instead. In contrast to parametric statistics, where statistical significance tests also consider characteristics of the population, the statistical significance of non-parametric tests is calculated exclusively based on the information contained in the sample. Examples of non-parametric methods applicable for variables measured at nominal or ordinal scales are Mann-Whitney U and Kruskal-Wallis for testing differences between groups and Spearman Rho and Chi-squared test for testing relationships between two variables [AT09].

2.7.3 Statistical Hypothesis Testing

As a method of inferential statistics, null hypothesis significance testing allows testing an experimental factor against a defined hypothesis. The test is based on the formulation of two hypotheses: a null hypothesis H_0 and an alternative hypothesis H_1. While H_0 formulates no effect of the experimental factor, H_1 formulates an effect of the experimental factor [Per15]. The test result is a P-value representing the probability that the data

could have occurred, assuming the null hypothesis was true. A low P-value indicates low support for the null hypothesis and thus indicates the validity of the alternative hypothesis. Before performing the test, the significance level is set, which corresponds to the probability of rejecting the null hypothesis even though it is true. The result of the hypothesis significance test, the P-value, is compared with the defined significance level. If the P-value is less than or equal to the defined significance level, the null hypothesis is rejected in favour of the alternative hypothesis [BL07].

Testing for difference: The use of significant difference tests allows examining whether the measured difference in means between two or more groups is due to an actual difference in the underlying population or to chance emanating from the sample. A commonly used parametric method for testing the difference between two groups is the Student's t-test, a hypothesis test that defines a null hypothesis (no significant difference existing) and an alternative hypothesis (difference existing). If more than two groups need to be compared, an ANOVA (analysis of variances) can be used [Vog21]. Depending on the study design, both methods provide approaches to examine differences between groups of different subjects (independent measurements) or groups of the same subjects (repeated measurements). If the requirements for performing a parametric test are not met, the corresponding non-parametric tests can be used. Non-parametric alternatives for the Student's t-test are the Mann–Whitney U test for independent measures and the Wilcoxon signed rank test for repeated measures. Non-parametric alternatives for ANOVA are the Kruskal–Wallis test for independent measures and the Friedman test for repeated measures [BSW14].

Testing for relationships: In addition to testing for differences in means between groups, statistical significance tests can be used to examine the significance and strength of a relationship between two variables. A common parametric statistic to describe the linear correlation between two interval or ratio scaled variables is the Pearson product-moment correlation (Pearson's r). The resulting Pearson's r describes the strength and direction of the relationship, ranging from -1.0 (perfect negative relationship) to +1.0 (perfect positive relationship). Similar to testing for differences, a relationship's significance is determined by the resulting P-value, which is compared to the defined significance level. Non-parametric alternatives are the Spearman rho correlation coefficient and the Chi-squared test, which can be used to examine the relationship between two variables at the ordinal and nominal levels of measurement, respectively [AT09].

2.8 Related Work

Existing learning systems in martial arts mainly focus on teaching techniques and improving rule knowledge. While applications like *Fight Trainer* [fig] and *Martial Codex* [mar] aim to improve athletes' techniques, the platform *WAKO Rules Live* [wakb] teaches rules for different kickboxing disciplines by rendering 3D animations. Apart from the *WAKO Rules Live* platform, which was no longer available at the time this study was conducted, there is no training program focused on developing the skills of martial arts

Table 2.3: Characteristics of reviewed video-based decision-making training programs.

Study	**[GY16]**	**[SPKB11]**	**[LMB$^+$18]**	**[MCMM05]**
Sport	Soccer	Soccer	Football	Rugby
Scope	Rules Decisions	Decisions	Decisions	Decisions
Content	Text-based, Video-based	Video-based	Video-based	Video-based
Approach	Game-based learning	Multiple-cue prob. learning	Implicit learning	Naturalistic learning
Feedback	Immediate	Immediate	No feedback	Delayed
Foundations	/	Social-cognitive model	Cognitive elim. process	Shared mental models

referees. However, there are training programs for referees from other sports, some of which have been scientifically designed and evaluated for effectiveness.

The following sections provide an overview of selected platforms aiming to foster referees' rule knowledge and decision-making skills in various sports. In particular, the studies on video-based decision-making training programs provide an important knowledge base for designing and evaluating similar systems.

2.8.1 Studies of Video-based Decision-Making Training for Referees

Although no studies were conducted to improve the decision-making skills of referees in martial arts explicitly, studies in other sports evidenced the effectiveness of utilising video-based platforms to improve referees' decision-making skills. Many studies on decision-making platforms for referees emphasise the potential of video-based programs as a complementary training method. According to skill development frameworks, like the 10.000h rule of deliberate practice [EKTR93], a high magnitude of training intensity needs to be accumulated to reach expert level in certain skills. Therefore, researchers argue that using video-based programs can help referees reach expert levels in decision-making by enabling a training intensity that is hardly achievable through participation in real-life competitions alone.

The studies discussed in this section examine training programs that aim to improve decision-making skills for referees. While all studies are based on platforms utilising videos as training material, they attempt to achieve the goals by different approaches. Table 2.3 outlines the main characteristics of these training programs.

Serious game for soccer referees

In a study conducted by Gulec and Yilmaz [GY16], a serious game was developed to improve the learning experience and decision-making competencies of soccer referees in Turkey. The platform was designed according to a digital board game, with the game elements of bonus fields and a leaderboard based on racecourse completion time. The platform covered three kinds of questions presented to users after rolling a virtual dice: true/false questions, multiple choice questions, and video questions. While the true/false and multiple choice questions were selected from a standardised pool of rule knowledge questions, the video questions included selected video scenes depicting certain situations to be judged. Players could only move forward if they correctly answered the question. Figure 2.12 illustrates the game board of the serious game.

The development of the serious game was motivated by the disadvantages of solely learning based on the International Federation of Association Football's (FIFA) Laws of the Game (LOG) book, which consists of textual information conveying the game's rules. The authors of the study argue that this is no suitable learning material to improve decision-making skills of referees. To address these shortcomings, the training program examined in this study aims to improve referees' rule knowledge and decision-making skills with the help of multimedia content. The design of the serious game was grounded on the positive results of studies utilising game elements to improve the learning experience. Gulec and Yilmaz claim that it was the first study to improve soccer referees' knowledge level and decision-making skills by utilising game design techniques.

The serious game was evaluated by applying both quantitative and qualitative methods. The results showed that the serious game approach was more effective than the traditional learning approach with the official FIFA LOG book, and a higher training intensity positively affected the score increase. In addition, qualitative feedback suggests that the platform has the potential to increase the self-confidence of referees, and even simple game elements can contribute to engaging referees in the training platform. The platform was perceived as a beneficial training tool, especially for exam preparation. Gulec and Yilmaz conclude that the training platform can be a beneficial tool to educate referees about the game's rules. However, positive effects on decision-making competencies were only implied and not explicitly examined in the study. Although solid rule knowledge is a prerequisite for making proper decisions [PH06], the training platform cannot be classified as a tool to improve referees' decision-making skills in the narrower sense.

Video-based training program for soccer referees

Compared to the serious game to convey rule knowledge introduced by Gulec and Yilmaz [GY16], Schweizer et al. [SPKB11] examined a video-based training platform to improve intuitive decision-making skills of soccer referees. Brand et al. [BPS09] describe the functionality and conceptual considerations behind the development of the training platform called SET (Schiedsrichter-Entscheidungs-Training, in English: referee decision-making training). SET presents video clips of potential foul situations to referees, who

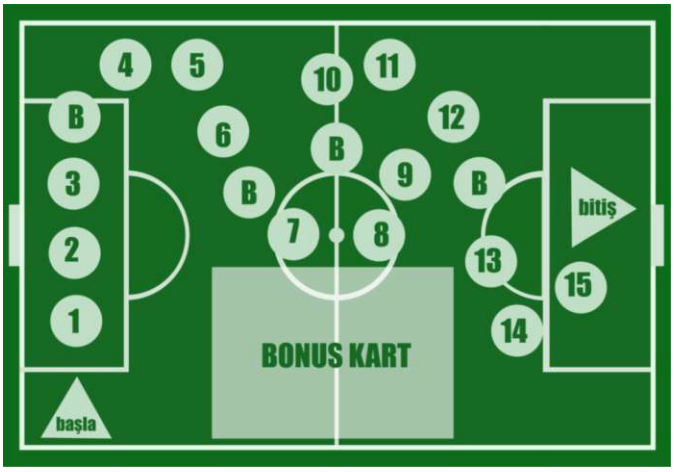


Figure 2.12: Game board of the serious game to improve the learning experience for soccer referees [GY16].

must decide whether the scene posed a foul or no-foul situation under time pressure. If the decision *foul* was selected, an appropriate sanction must be chosen (free kick, yellow card, red card). Subsequently, feedback on the correctness of the decision is provided. Figure 2.13 shows the screen prompting the referee to decide whether or not a foul occurred. The inputs resulting from the judgement (decisions and reaction time) are persisted, which enables the evaluation of the referees' performance. In addition to the training module, an administration interface provides methods to upload video scenes and annotate them with decisions at specific time points to provide appropriate feedback by comparing the user inputs with the defined correct decisions. The prepared material can be compiled into training sessions with different emphasises and released to certain groups of referees. The mode of feedback can be configured with respect to the extent of feedback displayed (only correctness vs correctness and complementary explanation), latency (immediate vs delayed) and aggregation (feedback for each decision vs feedback for several decisions in bulk).

The training program was designed based on the social-cognitive model introduced by Bless et al. [BFS04], who state that decisions are derived by traversing a sequence of social information processing steps, including perception, categorisation, memory processing and information integration. While all four steps are essential to derive a proper decision, each step's emphasis depends on the characteristics of the judged situation. On the one hand, the detection of offside situations requires intense demands in the step of perception, as the position of the attacker, defender and ball needs to be taken into account. On the other hand, detecting potential foul situations requires more emphasis in the categorisation step, as an accurate perception does not guarantee the derivation of an accurate decision. Therefore, SET focuses on the step of categorisation, while stating that improvements in this phase will also benefit subsequent steps. As a learning paradigm to train the categorisation of perceived information, the principles of

multiple-cue probability learning [LNKS06] and Hogarth's approach [Hog11] for learning intuitions are applied. More information on these principles can be found in sections 2.1.3 and 2.1.4.

The video-based training platform SET was examined through two field experiments, each conducted as a *pretest-posttest control group design*. The results showed that immediate feedback led to better scores than delayed feedback provided at the end of the training session. Moreover, extending the feedback with an additional repetition of the video scene was not more effective than immediate feedback alone. These findings suggest that immediate and non-repetitive feedback indicating the correctness of the judgement is sufficient to improve intuitive decision-making skills of referees.

Figure 2.13: Decision screen in the decision-making training tool SET [SPKB11].

Video-based training program for Australian football referees

Larkin et al. [LMB+18] evaluated a video-based training approach to improve the perceptual-cognitive decision-making process of Australian football referees. The training program was not based on an interactive digital application but on a compilation of video scenes depicting potential infringements. After watching the video scene, the referees were prompted to document their decision together with the justification in an answer booklet within a maximum period of six seconds after the occurring incident. The training material consisted of four different types of videos: (i) *full event viewing* videos including the complete decision-making scene stopping one second after the incident, (ii) *at event viewing* videos including the evolving situation stopping 0.5 seconds after the incident, (iii) *pre-event + outcome viewing* videos showing the same scene twice with different occlusion points 0.25 seconds prior the incident respectively after the decision indicated by the referee in the footage, (iv) *multiple clip viewing* videos consisting of a sequence of five short clips transitioning 0.5 seconds after the incident. No explicit instructions

in the form of pre-cues, information-rich areas or key teaching points were presented to the referees. Except for the videos of category (iii), which reveal the referee's decision in the second run of the scene, no feedback about the correct decision was shown to the referees. Thus, the only source of information was the set of compiled training videos.

The design of the training program was based on the principles of implicit learning and the importance of early identifying decision-making cues. Although explicit instruction can foster skill acquisition, it can negatively influence the retention of learning outcomes in stressful environments. Therefore, the training program aims to foster the retention of the learned skills by utilising an implicit learning approach, where feedback and instructions are reduced to a minimum. Another consideration integrated into the design of the training program was that referees derive decisions by applying a cognitive elimination process, where each decision leads to a further sub-decision until the final outcome is derived. The authors suggest that this process relies on information early in an action sequence. Considering these theoretical inputs, the program utilised video material for training the early identification of decision-making cues to support the elimination process.

The effectiveness of the training program was examined based on a *pretest-posttest control group design* with two retention tests to examine the development of training effects. Although the referees' results did not show significant improvements one week after conducting the training phase, their scores increased after a further two weeks. This observation might be explained by the assumption that implicit learning approaches require more time to acquire skills, and the consolidation of the learned skills was not completed after one week. The results also showed that experienced referees reached higher scores in the pre-test. However, while experienced referees did not significantly improve during the training period, the scores of less experienced referees converged over time. This phenomenon can be explained by the power law of practice [NR81], where the learning curve flattens over time as the practice continues. The authors conclude that the complementary use of video-based training programs can improve Australian football referees' decision-making skills.

Video-based training program for rugby referees

By emphasising the importance of properly reason the correct decision, Mascarenhas et al. [MCMM05] examined a video-based learning program aiming to improve the coherent decision-making skills of rugby referees. The training program, which is compiled of classical videotapes, is recognised as pioneering work in the field of video-based programs to improve decision-making skills for referees [KCL+21]. The referees' task was to watch and assess a series of videos showing various kinds of tackles recorded from a referee's perspective. At the beginning of each video scene, a voice-over introduced basic contextual information necessary to interpret the situation. The actual tackle incident started five seconds after the beginning of the video, allowing the referees to orient themselves in the scene. Immediately after the tackle, the video was frozen, and the referees were prompted

to make a decision within five seconds before the video turned into a blank screen. The referees were requested to indicate their decision, the reasons for that decision, and their confidence level. At the end of a series of video scenes, the videotape provided delayed feedback, including (i) a replay of the tackle, (ii) a detailed explanation by a highly-acknowledged referee, and (iii) another replay of the tackle, which can be watched repeatedly.

The design of the training program was shaped by the principles of *naturalistic decision-making* and *shared mental models*. Mascarenhas and Collins suggest that decisions should be trained in an environment of high ecological validity reflecting real-world constraints like stress and crowd noise, referred to as *naturalistic decision-making*. Knowledge needs to be organised in proper knowledge structures to rapidly derive decisions in such environments. The organisation of this knowledge can be represented in the form of *mental models* holding information required to derive an appropriate action. In the context of decision-making within a team, the term *shared mental model* is used to describe that team members have compatible knowledge structures.

Taking these considerations into account, Mascarenhas et al. refined the definition of *accurate decisions*. They empathise that it is vital to not only evaluate whether the decision is correct but also if the decision was adequately derived. Thus, a decision can only be considered accurate if both decision and reasoning are based on a complete mental model of the situation. If a group of referees derive their decisions based on a *shared mental model*, they can be described as *coherent*. The training program aims to improve referees' coherent decision-making processes in a natural decision-making environment.

The effectiveness of the training program was examined based on a *pretest-posttest control group design* with an intervention consisting of five training sessions distributed over six weeks. The experiment's results showed that the accuracy did not exceed the value of 60 %. However, it needs to be considered that accuracy is defined as the combination of both correct decision and correct reasoning. While lower-ranked referees significantly improved their accuracy from pre-test to post-test, mid- and high-ranked referees only showed minor, non-significant improvements, comparable to the results in [LMB+18]. Based on these findings, the training program was considered a suitable tool for developing shared mental models of pre-elite referees.

2.8.2 Commercial Referee Platforms

Apart from the studies examining video-based decision-making platforms described above, commercial platforms for referees are also available. Although their effectiveness is not examined in the course of studies, they might be beneficial to enhance referee-specific knowledge through digital techniques. However, these platforms' primary purpose is to facilitate rule knowledge. Although rule knowledge is a prerequisite to making proper decisions [PH06], these platforms cannot be classified as decision-making training programs in the narrower sense.

Get It Right

The platform *Get It Right* [get] aims to teach rules to referees, coaches and athletes by using an animation-based training approach. While it primarily focuses on the sport of American football, it also supports the sports of Volleyball and Wrestling.

The explanations are based on three-dimensional animations, which allow visualising scenarios from multiple perspectives. The rules module facilitates rule knowledge by rendering animated scenarios, highlighting information-rich areas and providing explanations according to the official rule book (see Figure 2.14a). The mechanics module shows animations of situations a referee is likely to encounter during a game and provides instructions on how they should act. The so-called spotlight feature indicates which viewing range the referee should cover in certain situations (see Figure 2.14b). In order to test the gathered rule knowledge, an evaluation module is provided, where selected scenarios are presented, which need to be assessed by the user. The assessment not only requires indicating the final decision but also covers rule knowledge and signals the referee has to indicate as a response to the presented scenario. At the end of the assessment, the percentage of correct answers is displayed, and the user can review them individually (see Figure 2.14c).

Although the platform's user interface is built with the help of game technologies, game elements are only used in the evaluation module's assessment feature. Based on the platform's spectrum of functionalities, it can rather be considered a rule-learning tool. Although rule knowledge is a prerequisite to making proper decisions [PH06], it cannot be classified as a tool to improve the decision-making skills of referees in the narrower sense.

Zooom App

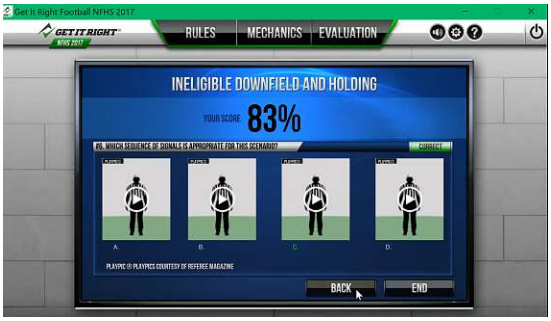
Zooom App [zoo] is a platform with the vision to improve the performance of referees by using videos as learning material. Its core functionality is to teach rules and allow users to easily create and share training videos with platform users. As it also aims to connect referees within the platform, it also has a strong social character.

The platform's functionality consists of multiple modules. The module *RefPrep Officiating E-Learning* is designated for referee students and instructors. It provides courses prepared by professional referees in American football, baseball, basketball, soccer and volleyball. It teaches philosophies, rules and mechanics based on reviewing augmented videos recorded from different angles. Other modules are designed for organisations and include functionalities for preparing the video content by providing utilities to highlight information-rich areas, add textual descriptions, add voice-overs and organise videos into libraries. The "You Make the Call" feature allows users to upload a video and ask the referee community for feedback.

Due to the lack of game elements, the rule-teaching platform cannot be categorised as a serious game. Therefore, although it might positively influence referees' decision-making

(a) Facilitation of rule knowledge.

(b) Basic scenario with spotlight visualisation.



(c) Selection of the correct referee gesture in the evaluation module.

Figure 2.14: Functionalities of the *Get It Right* platform [get].

skills by increasing rule knowledge, it cannot be classified as a platform to improve the decision-making skills of referees in a narrower sense.

2.9 Summary

The literature cited in this chapter provides the theoretical background for the analysis, design, development, and evaluation of a serious game to train the decision-making skills of martial arts referees. It covered the research of scientific work in the area of decision-making processes, video-based decision-making training programs, and serious games, as well as methods for constructing and evaluating prototypes like the one built in the course of this thesis. Therefore, the theory described in this chapter serves as a source of information for conducting the study, framing its scope, and answering the defined research questions defined in section 1.4.

CHAPTER 3

Methods

For the development and evaluation of the serious game, the comprehensive methodological approach of Design Science Research was used in this work. Hevner et al. [HMPR04] introduce a conceptual framework to understand, execute and evaluate design science research. With the help of seven guidelines, this framework aims to facilitate the requirements to conduct effective design science research. According to these guidelines, a central aspect is the iterative creation of an innovative and purposeful artefact to address yet unsolved and important problems in specific domains. In order to demonstrate the artefact's utility, it must be evaluated according to requirements defined by the business environment. The research must be conducted by applying rigorous methods in the construction as well as in the evaluation of the artefact. This is accomplished using the knowledge base of theoretical foundations and research methodologies. In return, the results of the design science research are published and thus contribute to expanding the cumulative knowledge base.

The subsequent sections describe the scientific methods applied in the context of the design science research conducted in this work. After outlining the general methodological approach, each applied method is described in separate sections.

3.1 Methodology

In the course of the design science research conducted in this work, different methods were applied to create and evaluate the artefacts as a prerequisite to answering the defined research questions. The research was based on a domain-specific *literature review*, which enabled the study to reach a high level of quality by taking findings from existing studies into account. This also contributes to the criteria of scientific relevance demanded by design science research, as it emphasises the knowledge gap in the concerned area [HMPR04]. Considering the inputs from the literature review, *requirements* defining the

functionality of the serious game were collected using qualitative methods. The *design and development* of the serious game was based on the method of prototyping. This enabled to rapidly create tangible artefacts allowing stakeholders to envision and reflect on the final system in early phases [BLM09]. Feedback and new requirements resulting from the iterative development cycles were incorporated accordingly. In the final step, the developed prototype was evaluated using the empirical methods of a field experiment and a questionnaire.

Figure 3.1 visualises the research process outlined above. For each research step, it shows the applied method as well as the produced outcome, which serves as the basis for answering the defined research questions. The following sections provide further insights into the methods applied in the course of the design science research applied in this work.

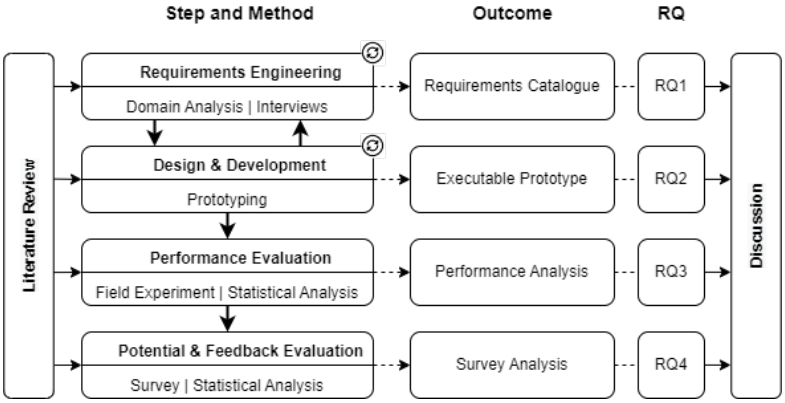


Figure 3.1: Phases of the work, including applied methods and resulting artefacts.

3.2 Literature Review

The literature review provided the theoretical background for conducting the present study. It served as input for all phases of the research process and contributed to the scientific rigour of this study. In particular, insights from the following domains were reviewed and used as input to this work.

Decision-Making Processes: In order to gain insights into different types of referees and their associated decision-making demands, literature in the context of decision-making processes was reviewed. The classification introduced for characterising referees according to their primary demands allowed to broaden the knowledge base by drawing on conclusions from other sports in which referees face challenges similar to those in martial arts. The research of literature on the social-cognitive procedure behind concluding a decision was pivotal for designing the serious game.

Video-based Decision-Making Programs: To learn from related studies, state-of-the-art literature on video-based decision-making programs for referees was reviewed. Since

no studies on decision-making programs for martial arts referees had been published at the time of this study, the scope of the literature review was extended to include studies on video-based decision-making programs for sports referees in general. Based on the decision scenarios that the examined program was supposed to train, it was deduced whether the conclusions were transferable to martial arts refereeing.

Serious Games: The decision-making program developed in the course of this thesis was not solely intended to be a training program but also a serious game. Therefore, literature in the field of serious games was reviewed to gain insights into design principles for serious games. This also allowed classifying this work's serious game into the sub-category of digital game-based learning, further refining the scope of the literature review.

Prototype Construction and Evaluation: Besides studying fundamental literature that influences the design and mechanics of the serious game, the literature review also encompassed the research on prototyping approaches as well as applicable methods to build and evaluate the prototype. In this way, not only is the mechanics of the prototype based on a solid knowledge base but the construction and evaluation of the prototype are also performed using scientifically sound methods.

3.3 Analysis, Design, and Development

The development of the prototype encompassed activities of requirements engineering, design, and development. The following paragraphs describe the methods and communication structures used in these phases.

3.3.1 Requirements Engineering

Building upon the fundamental knowledge base and state-of-the-art literature research, initial requirements of the serious game were gathered through domain analysis [ZC05] based on already existing artefacts. This included the analysis of (i) an already existing primitive video-based prototype developed for the sake of showcasing the idea to stakeholders and (ii) rudimentary functional and technical concepts. Although this source of requirements did not represent a sufficiently detailed basis for the development, it provided a solid basis for discussion to collect further details through subsequent expert interviews.

Based on the findings from the domain analysis, semi-structured interviews with subject matter experts [NHWA15] in kickboxing and karate Kumite were conducted. Since the system's main functionalities were already gathered at this point, the scope of the interviews was limited to the refinement of existing and identification of missing requirements. The interviews were structured in (i) an introductory section explaining the idea of the system to be developed, (ii) the presentation of the planned system encompassing its main functionalities and boundaries, and (iii) a question round to clarify

and refine ambiguous requirements. All expert interviews were solely conducted virtually by using *Microsoft Teams* [ms] and *Jitsi Meet* [jit].

3.3.2 Design and Development

Based on initial gathered requirements, the serious game was designed and developed through the method of evolutionary prototyping [Flo84]. This phase was shaped by building artefacts, followed by retrieving and incorporating feedback from field experts. After two iterations of exploratory prototyping, the final prototype was built through evolutionary system development in seven iterations.

Exploratory prototyping: Based on initially gathered requirements, exploratory prototyping encompassed the design, demonstration, evaluation, and refinement of low-fidelity prototypes. During this phase, the prototype was used as a vehicle allowing participants to envision the structure and functionality of the target system. Questions about ambiguous features were addressed to the audience and discussed directly in the context of the respective screen. The participants were encouraged to actively give feedback and ask questions relevant to the current context. The feedback from these user sessions was used to (i) gradually refine the low-fidelity prototype and (ii) improve the quality of the requirements catalogue.

Evolutionary system development: The method of evolutionary system development resulted in executable prototypes demonstrated to subject matter experts in kickboxing and karate Kumite. Feedback from these sessions was used to gradually refine the prototype until all requirements were met and the prototype reached maturity for field testing.

Experimental prototyping: During evolutionary system development, experimental prototyping [Flo84] was selectively applied to evaluate the feasibility and performance of certain aspects of the serious game. Rather than basing system development on uncertain assumptions, proofs of concept were created to test the suitability of certain technologies and libraries for implementing specific functionality in early development phases. In addition, the performance of resource-intensive features was tested under realistic conditions by filling the database with the expected data volume.

3.3.3 Work Allocation and Communication

For development-related tasks, a collaboration was organised with another student to deliver the gaming interface of the serious game. The work split was determined based on the gathered requirements by considering (i) the expected workload, (ii) the adhesion of features, and (iii) the limitation of dependencies. To enable efficient communication, a daily virtual Microsoft Teams [ms] meeting was scheduled to exchange experiences, discuss potential problems, and coordinate the integration of produced artefacts.

Work Breakdown: Requirements were divided into epics and user stories and documented in the DevOps platform Gitlab [git], which made it possible to assign tasks transparently and track the progress of feature development. The decomposition of work into small tasks and a resulting reduced lead time enabled an incremental development process [KS10]. This allowed early and recurring feedback within the development team based on the frequent deployment of features.

3.4 Field Experiment

To assess the decision accuracy and reaction time of licensed kickboxing referees in the serious game, a field experiment was conducted during the Austrian WAKO Kickboxing Championship 2022 in Graz. This section describes the structure of the field experiment, including its procedure, the characteristics of the participants and the material used.

3.4.1 Participants

The sample consisted of 16 licensed 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). While all participants were officially licensed referees for the Tatami disciplines of Point fighting, Light contact and Kick light, only 6 had licenses for the Ring disciplines of Full contact, Low kick and K1 Style. 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). The referees' age and average refereeing experience were calculated based on a survey completed by 14 of the 16 referees participating in the field trial.

All persons voluntarily participated in the study and indicated their agreement by signing a consent form. Ethical approval of the study procedure was obtained by the TU Vienna Ethics Committee.

3.4.2 Procedure

The field experiment followed a procedure consisting of four activities completed in two blocks distributed over two days. To inform participants about this procedure and the basic functionality of JudgED, an explanatory video was produced and shared with participants one week in advance. In addition, before starting the field experiment, the participants were informed about the scope of the study and instructed about the upcoming activities.

The following paragraph outlines the steps of the field experiments in chronological order.

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 the videos V_{T1}, the Ring referees S_R judged the videos V_{R1} in the serious game.

Day 2:

1. Test 2: While the Tatami Referees S_T judged the videos V_{T2}, the Ring referees S_R judged the videos V_{R2} in the serious game.
2. Reaction Time Test: All participants completed the standardized *four choice reaction time test* [fou].

This procedure was conducted in a physical room equipped with chairs and a sufficiently strong internet connection. All activities were performed on 10-inch Android-based tablets to provide a screen size suitable for detecting the athletes' movements in the video scenes. By describing the tasks the participants were requested to perform, the following two sections provide detailed information about the activities included in the field experiment.

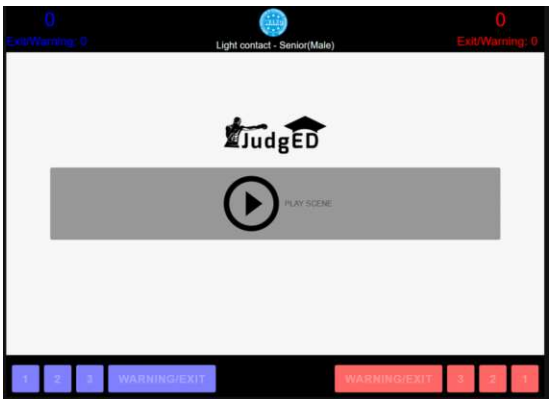
Familiarization, Test 1 and Test 2

Activities 1-3 of the field experiment were carried out on the mobile Android version of the serious game. The participants' task in these activities was to judge the video scenes in the playlists as accurately and fast as possible. For each participant, separate login credentials were created. In order to use the available time for the field experiment efficiently, the tablets were already configured with registered users before they were handed over to the respective participants.

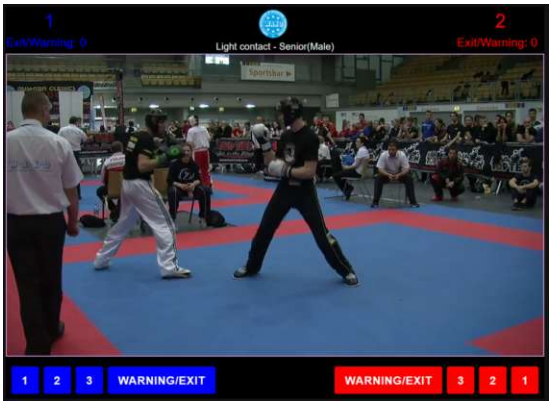
Figure 3.2 depicts the different steps of the judgement process executed by the participants. Each video scene was started by clicking on the start button (Figure 3.2a). While the video was in progress, the participants indicated their judgments for the respective athlete on the touchscreen (Figure 3.2b). Referees visible in the original footage were blurred in case they indicated a revealing gesture (Figure 3.2c).

Four Choice Reaction Time Test

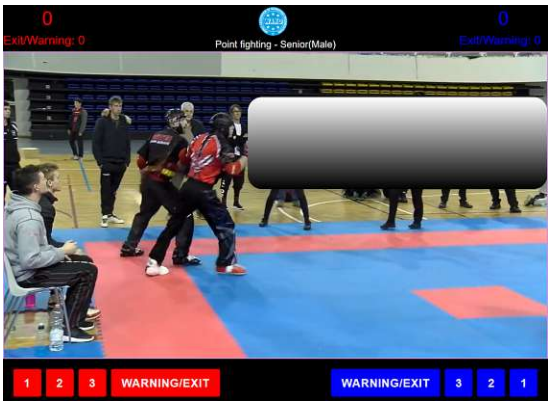
In the last step of the field experiment, the participants were requested to perform a reaction time test on the 10-inch tablets provided. The mobile version of the Inquisit Player application [inq] was used to run a *four choice reaction time test* [fou]. Figure 3.3 shows the screen displayed to the participants during this test. Four boxes were visible at the top of the screen, only one of which was randomly highlighted in each test round. Four buttons appeared in the lower part of the screen. The participants' task was to press the button corresponding to the position of the highlighted box as quickly and accurately



(a) Start video.



(b) Video in progress.



(c) Blurred referee after a decision in Point fighting.

Figure 3.2: Screens of judging video scenes, feedback and personal dashboard.

as possible. After pressing the button, a new round was automatically started, again randomly highlighting a box. The test consisted of 10 rounds, which allowed participants to familiarise themselves with the task before starting the actual test, which consisted of 100 rounds.

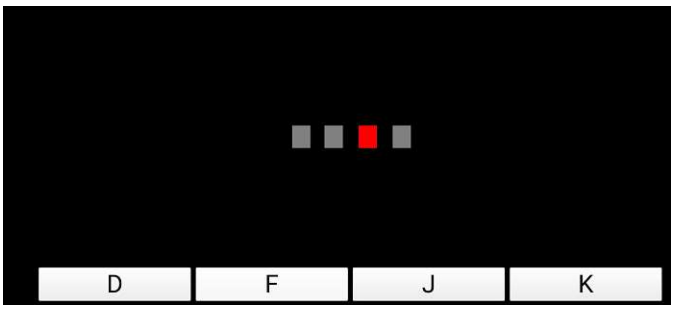


Figure 3.3: Four-choice reaction time test [fou] in Inquisit Player [inq].

3.4.3 Material and Setup

Activities 1-3 of the field experiment were conducted in JudgED, which relies on video scenes to be selected, uploaded, and prepared for the use in the serious game. This section describes specifics about the source, preparation and compilation of video scenes, and its setup in JudgeED.

Source of Footage and Video Scene Creation

Except for three video scenes recorded from private training sessions, the video scenes used for the field experiment consisted of kickboxing footage from real-world competitions produced by Brannmanndan[1]. Permission to use these videos as a source for the serious game was obtained from the originator. This enabled the selection of the appropriate footage from tens of thousands of videos. All videos had a minimum resolution of 720p.

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 $>$ 10 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 has also experience as a coach (13 years) and scientist (7 years). The video scenes were only annotated with decisions in case the respective action was clearly 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, 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 organisation process was performed exclusively in JudgeED through functionalities of the content and administration module.

Sets of Video Scenes

For activities 1-3 of the field experiment, distinct sets of video scenes were prepared by considering the participants' referee licenses. 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. Table 3.1 summarises the characteristics of these sets with respect to the number of included video scenes, the net duration, the athletes' gender, and the mean difficulty.

The video sets V_{T1}, V_{T2}, V_{R1} and V_{R2} were compiled in such a way that the throughput time does not exceed 15 minutes and the average level of difficulty between Tatami and

[1] https://www.brannmanndan.com

Table 3.1: Characteristics of the video scenes for Tatami and Ring disciplines.

Playlist	# Video Scenes	Duration	Male/Female	Mean Difficulty
$\mathbf{V_{Tatami\text{-}1}}$	25	5:25 m	24/1	2.6
$\mathbf{V_{Tatami\text{-}2}}$	27	6:32 m	23/4	2.6
$\mathbf{V_{Ring\text{-}1}}$	12	11:56 m	11/1	3.1
$\mathbf{V_{Ring\text{-}2}}$	12	12:40 m	6/6	3.1

Ring video sets is equal. The lower total net duration of the Tatami video sets is caused by the fact that video scenes of the Tatami disciplines are shorter and thus require more time to transition between them.

JudgED Setup

The videos sets V_{T1}, V_{T2}, V_{R1} and V_{R2} were organized in form of playlists in JudgED. This allowed sharing them with the participants of the field experiment. All playlists were configured with mode *Exam*, which caused included video scenes to appear in their defined order. In addition, the playlists were not repeatable, and no feedback was shown to the player after judging the video scene.

For each participant, a user was created in JudgED that allowed them to access the designated playlists for the period the field experiment took place. In order to protect the privacy of the participants, no real names but pseudonyms were used as usernames. Only the author of this work had access to the list linking the pseudonyms to the real names.

3.4.4 Data Processing

The metrics to be examined by the field experiment are the decision accuracy and reaction time of the participating referees. Therefore, the serious game computed and logged each decision's correctness and reaction time in the JudgED MongoDB database. However, the reaction time is only available for entered decisions that could have been matched with defined decisions. The logged data for test 1 and test 2 was queried by MQL queries and exported as CSV files for further processing.

The results of the four-choice reaction time test [fou] were logged by the Inquisit Player mobile app [inq] in two separate CSV files per participant: (a) a summary file with the participant's average result and (b) a detail file with the results for each of the 100 test rounds. For this study, only the summary file was used to determine each participant's overall reaction time.

3.5 Survey

In order to receive additional feedback from the participants of the field experiment, a survey was conducted. The purpose of the questionnaire was to gather (i) basic demographic data, experience, and opinions about computer-supported referee education, (ii) feedback on certain topics relevant for future feature development, (iii) the self-assessment of decision-making abilities, and (iv) opinions about potential application areas and target groups of the serious game. The following sections provide information on the design, administration, participants, characteristics, and analysis of the survey.

3.5.1 Questionnaire Design Process

The questionnaire design was based on the iterative process described by Sheatsley [She83]. After the purpose of the questionnaire was defined, the questions were drafted and placed in a meaningful order before being reviewed by two subject matter experts in multiple iterations. The group of reviewers included two subject matter experts: (i) an official WAKO kickboxing referee (experience $>$ 10 years) and (ii) a professional WAKO kickboxing athlete (experience 16 years) who has also experience as a coach (13 years) and scientist (7 years). While the first versions of the questionnaire were created as a plain text document, subsequent versions were already created on the platform used to administer the survey. The final questionnaire is available in Appendix 7.

3.5.2 Participants and Administration

The questionnaire was sent out to the group of referees participating in the field experiment. Since the sample consisted exclusively of German-speaking participants, the questionnaire and all accompanying instructions were created in German. Section 3.4.1 provides a detailed summary of the participants' characteristics. The questionnaire was completed by 14 of the 16 referees.

The questionnaire was created on the platform SurveyMonkey[2] and conducted as a self-administered online survey. The participants were informed about the questionnaire via email and WhatsApp[3]. The invitation to the questionnaire included a participant-specific link to the questionnaire and basic instructions on how to get started. The questionnaire was made available to the participants two days after the field experiment had been carried out.

In order to protect the privacy of the participants, no real names but pseudonyms were used to manage them in SurveyMonkey. A unique questionnaire URL was generated for each respondent, allowing the responses to be linked to the participants' pseudonyms. Only the author of this work had access to the list linking the pseudonyms to the real names. This approach made it possible to correlate the questionnaire responses with

[2] https://www.surveymonkey.de/
[3] https://www.whatsapp.com/

the results of the field experiment. All participants voluntarily participated in the study and indicated their agreement by signing a consent form. Ethical approval of the study procedure was obtained by the TU Vienna Ethics Committee.

3.5.3 Question Types, Response Categories & Quality Criteria

As the questionnaire aims to gather information about certain opinions in a pre-defined frame rather than generating new ideas, the format of closed questions was chosen with a few exceptions. This approach corresponds to the general recommendation of Sheatsley [She83] to use closed questions wherever possible.

Depending on the kind of question, different response category scales were applied. While nominal scales [FM09] were used to capture specific qualitative data, ordinal scales [GMR07] were used to measure the participants' tendencies. Dichotomic scales were used to capture responses to questions where no nuances were applicable. To measure the respondents' attitude towards defined statements, 5-point Likert scales with fully-labelled response categories were used [She83, WCS10, CRSH17].

To increase the quality of the questionnaire, questions were phrased by adhering to best practices defined in [She83]. Questions were kept short within a length of 25 words wherever possible. In addition, long and ambiguous words, technical terms, complex phrases, double-barreled questions, false premises and double negatives were avoided. To further increase comprehensibility, some questions were supported by complementary multimedia content in both question and response items.

3.5.4 Data Processing

Each participant's responses to the respective questions were logged by the platform used for administering the online survey. The survey platform's functions were used to retrieve the responses (i) in an aggregated way to summarize the results and (ii) for each participant separately to combine it with data from the field experiment.

The two positive (agree, totally agree) and negative (disagree, totally disagree) response items on questions with a 5-point Likert scale were combined to report the results. The questions F2, F3, F5, F6, F8, F9, F10, F11, F12, F13, F16, F21, F22, F33, F40, F44, and F46 were excluded from the evaluation due to their irrelevance to answer the research questions.

3.6 Statistical Analysis

The data recorded during the field experiment (section 3.4.4) and the survey (section 3.5.4) was analysed using descriptive and inferential statistical methods.

The data collected through the field experiment was analysed with the help of the software IBM SPSS Statistics [sps]. The Fleiss' kappa inter-rater reliability test was performed

to measure the level of agreement among the referees on the judgment of each defined decision. For categorizing the results in *poor*, *moderate*, *good*, and *excellent* agreement, the guideline introduced by Koo and Li [KL16] was used. The binomial distribution formula was used to calculate the probability of deriving a correct decision in a team of three independent judges requiring a majority decision. To test the data for differences between test 1 and test 2, a Student's t-test for dependent groups was performed. The Pearson correlation coefficient was computed to test for the linear relationship between (i) the performance of referees and their refereeing experience in years and (ii) the referees' reaction time in JudgED and the four-choice reaction time test. The Spearman's ρ was calculated to test for the linear relationship between the performance of referees and the difficulty rank of judged video scenes. A Shapiro-Wilk test was performed to test for the normality of the data. All inferential hypothesis tests were performed at significance level $\alpha = 0.05$.

To analyse the data collected through the survey, descriptive statistical methods were applied. The arithmetic means was used to measure central tendency. Responses for demographic questions were used as a basis to perform inferential statistical methods to test for correlations with performance data collected through the field experiment.

CHAPTER 4

Results

This chapter presents the developed artefacts of the design science research encompassing the design and development of a prototype to train decision-making processes of martial arts referees. After outlining the iterative development leading to the final version of the prototype, the artefacts resulting from the analysis, design and development of the prototype are described: (1) stakeholders and requirements catalogue, (2) design mockups, (3) functional, architectural, and technological description of the final prototype, and (4) performance measurement algorithm.

Section 4.2 presents the identified stakeholders as well as the requirements catalogue, including the final set of requirements that have been gradually refined in multiple iterations of exploratory prototyping and evolutionary system development. Based on initially gathered requirements, exploratory prototyping resulted in a set of design mockups that were used as a communication vehicle to facilitate a shared understanding in early research phases before starting iterative development of the executable prototype (section 4.3). A functional description of the final prototype, as well as its architecture and applied technologies, are described in section 4.4 and section 4.5, respectively. In order to precisely measure the decision accuracy and reaction time of the players in the serious game, an algorithm was developed, which is described in section 4.6. Some of the results presented in this chapter have already been published in a conference paper [SGFH+22].

4.1 Iterative Analysis, Design and Development

The content and administration module and the training module were developed in ten iterations encompassing analysis, design, and development activities. During these iterations, the persons described in Table 4.1 were involved as information sources to gather requirements and to collect feedback on developed artefacts. The iterations outlined

in Figure 4.1 visualize the development process from gathering initial requirements to developing the final version of the prototype, including its intermediate results. For each iteration, the diagram lists the developed features, associated requirements, and involved persons. Although all features have been refined over several iterations, the figure only lists them in the iteration in which they were initially implemented.

Table 4.1: Involved persons in the analysis, design, and development process.

ID	Description	Disciplines	Experience
P1	WAKO kickboxing referee	Tatami, Ring	> 10 years
P2	WAKO kickboxing referee	Tatami, Ring	4
P3	WAKO kickboxing referee	Tatami	> 10 years
P4	Former WAKO kickboxing athlete	Tatami	16
P5	Former WKF karate athlete	Kumite	10 years
P6	WAKO board member	/	/

Figure 4.1: Prototyping iterations including implemented requirements and involved subject matter experts.

4.1.1 Iteration 0: Requirements Engineering

The conduction of domain analysis and four semi-structured interviews with subject matter experts in kickboxing and karate Kumite resulted in the first version of the requirements catalogue, successively refined in further iterations. During the interviews, stakeholders crucial to the research project's success were also identified. The final version of the requirements catalogue and the identified stakeholders are described in section 4.2.

4.1.2 Iteration 1-2: Exploratory Prototyping

With the knowledge gathered in the preceding phase, these iterations resulted in low-fidelity prototypes encompassing design mockups for 20 screens that were rudimentarily linked together to generate the impression of a navigational flow. Uncertainties and missing information were filled with assumptions to be clarified during the prototype demonstration. The refined version of the low-fidelity prototype and the refined requirements formed the basis to start building the executable target system. The design mockups resulting from *Iteration 2* are described in section 4.3.

4.1.3 Iteration 3-9: Evolutionary System Development

Based on the preliminary design and refined requirements, the final version of the executable prototype was developed in seven iterations (iterations 3-9) through evolutionary system development. The final prototype is documented in the form of a functional description in section 4.4. A technical description of the prototype, including an overview of the data structure, architecture, and applied technologies, can be found in section 4.5.

4.2 Stakeholders and Requirements Catalogue

The two subsequent sections present (i) identified key stakeholders that served as a source of information throughout the requirements engineering and prototype design process and (ii) the final version of the requirements catalogue resulting from domain analysis, semi-structured, and evolutionary system development.

4.2.1 Stakeholders

The identified stakeholders contributed to the project's success by providing the necessary resources, promoting the project, and supporting the organisation of the prototype's evaluation. The following paragraph describes identified key stakeholders by highlighting their characteristics as well as their contribution and interest in the project.

Referee Coaches: Individuals with a wide range of refereeing expertise who also impart referee knowledge in the form of courses and seminars. They provide valuable feedback on the activities of requirements engineering, design, and development of the entire serious game comprising the training and content creation modules. Because they have a large network of referees, they help recruit referees needed to evaluate the prototype. Their interest is using the serious game as a complementary tool to enhance their conducted referee courses.

Referees: Licensed referees who have acquired a high level of knowledge throughout their careers as referees and as experts in the sport. Their refereeing knowledge provides valuable feedback for designing the training module. In addition, they are

involved in evaluating the prototype and are interested in using it to train their decision-making skills.

Content Creators: People with high refereeing experience and at least moderate computer skills for uploading videos and properly annotating them with decisions. They define video scenes that serve as the basis for evaluating the prototype. Their experience from the process of defining video scenes provides valuable feedback to improve the usability of the content creation functionalities. As this group of stakeholders overlaps with either referee coaches or referees, they have a similar interest in the project.

Content Providers: Persons producing and publishing videos of kickboxing competitions. Although they do not actively participate in the design process, they provide the footage on which the video scenes used in the serious game are defined.

WAKO: The World Association of Kickboxing Organizations (WAKO) organises various kickboxing competitions around the globe. The involvement of representatives of the WAKO referee committee for Tatami and Ring sports in the early stages of the project ensures their acceptance and support, as well as permission to consult WAKO referees throughout the design and evaluation of the serious game.

4.2.2 Requirements Catalogue

The information gained during the requirements engineering process disclosed the need to develop a serious game comprising two modules: (a) a training module to train the decision-making skills of referees by following a video-based training approach and (b) a content and administration module enabling authorised referees to define, organise and evaluate the training sessions.

Table 4.2 enumerates the identified high-level requirements classified in the classes of content and administration (C), training (T), and cross-cutting concerns (X). The table's origin column describes the iteration in which the requirement was identified. However, the requirements were gradually refined based on the feedback gathered during the design phase. For tracking purposes and to further specify these requirements as a basis for the development, they were further specified and broken down into epics and user stories.

In order to avoid redundancy between the descriptions of the requirements and implemented features, requirements are only described at a high level in this section. Instead, the sections 4.4.1 and 4.4.2 describe the functions of the developed prototype by referring to the related requirements they implement.

Content and Administration Requirements

C_1 **Upload videos:** The content and administration module shall allow administrative users to upload fight videos in MP4 format along with basic metadata describing

Table 4.2: High-level requirements classified in content and administration (C_i), training (T_i), and cross-cutting concerns (X_i).

ID	Description	Origin
C_1	Upload videos	0
C_2	Define and cut video scenes	2
C_3	Annotate video scenes with decisions	0
C_4	Highlight information-rich areas	0
C_5	Blur revealing referees gestures	2
C_6	Video scene status management	3
C_7	Compile video scenes in playlists	1
C_8	Configure feedback and playback modes	2
C_9	Release playlists for certain players	1
C_{10}	Release playlists for all players	6
C_{11}	Performance monitoring dashboard	0
C_{12}	Statistical performance evaluation	0
C_{13}	Manage users and permissions	1
T_1	Judge video scenes	0
T_2	Feedback on decision-level	0
T_3	Feedback summary on scene-level	0
T_4	Challenge defined decisions	1
T_5	Personal performance dashboard	0
T_6	User performance comparison	0
T_7	Mobile app support	0
T_8	Decision Data Persistence	0
T_9	Decision Matching	0
X_1	Authentication methods	0
X_2	Moodle integration	0
X_3	Extendability for other sports	0

the presented fight's characteristics. Metadata includes association, discipline, bout, tournament, year, age group, athlete 1, athlete 2, and gender.

C_2 **Define and cut video scenes:** The content and administration module shall allow administrative users to cut uploaded videos into multiple scenes of appropriate duration and describe them with basic metadata, including difficulty, main focus, and position of the red/blue athlete at the beginning of the selected time range.

C_3 **Annotate video scenes with decisions:** The content and administration module shall allow administrative users to annotate and extract video scenes with decisions according to the events occurring in the fight scene. Each decision is identified by the time, athlete, technique, and decision value. These configurations serve as a reference value to determine the correctness of the decisions entered by the players in the serious game.

C_4 **Highlight information-rich areas:** The content and administration module shall allow administrative users to highlight information-rich areas significant for detecting the correct decision. The highlighting shall be displayed as an overlay in the decision-specific slow-motion feedback presented to the player in the serious game for a defined period.

C_5 **Blur revealing referees gestures:** The content and administration module shall allow administrative users to blur referees visible in the video scene in case they indicate a decision. The blurring definition shall be displayed as an overlay in the video scene presented to the player in the serious game for a defined period.

C_6 **Video scene status management:** The content and administration module shall allow administrative users to manage the review cycle of video scenes in order to allow tracking their quality.

C_7 **Compile video scenes in playlists:** The content and administration module shall allow administrative users to compile video scenes into containers (i.e. "playlists") according to didactic or organisational requirements.

C_8 **Configure feedback and playback modes:** The content and administration module shall allow administrative users to configure different playlist modes in terms of allowed repetitions, playback order, and the extent of feedback displayed. Table 4.3 shows the three types of configurations determined by the modes *regular*, *lab*, and *exam*.

C_9 **Release playlists for players:** The content and administration module shall allow administrative users to temporarily release selected playlists to specific players of the serious game.

C_{10} **Release playlists for all players**: The content and administration module shall allow administrative users to release selected playlists to all players of certain disciplines.

C_{11} **Performance monitoring dashboard:** The content and administration module shall allow administrative users to display a dashboard that presents aggregated performance data from users in eligible courses, including decision accuracy, reaction time, and training intensity metrics. The dashboard shall also display video scenes with the lowest judgment performance in terms of decision accuracy and response time.

C_{12} **Statistical performance evaluation:** The content and administration module shall allow administrative users to generate charts for decision accuracy, reaction time, and training intensity on available fields of video scenes, users, and defined decisions.

C_{13} **Manage users and permissions:** The content and administration module shall allow administrative users to create and manage both administrative users

Table 4.3: Characteristics of regular, lab, and exam playlist.

Mode	Repetitions	Playback Order	Feedback	Slow Motion
Regular	Infinite	Random	Complete	Repetitive
Lab	Infinite	Random	Complete	Non-repetitive
Exam	One	Sequential	Not available	Not available

and players of the serious game. To manage the users' permissions, a role-based authorisation model shall be applied that supports the segregation of data based on the users' assigned associations.

Serious Game Requirements

T_1 **Judge video scenes:** The training module shall allow players to watch their assigned playlists' video scenes and to attribute decisions to the blue or the red athlete by using a discipline-specific scoreboard while the video is progressing.

T_2 **Feedback on decision-level:** The training module shall allow players to receive immediate feedback on the correctness of the player's judgments for each decision defined on the video scene. The feedback shall include the following information for each defined decision: (i) the player's decision, (ii) the correct decision, (iii) the reaction time, (iv) the correctness of the decision, (v) the applied technique causing the decision, (vi) a 30 % slow-motion video sequence spanning the 0.5 seconds period before and after the time of the defined decision.

T_3 **Feedback summary on scene-level:** The training module shall allow players to receive holistic feedback about their performance of the currently judged video scene, which also considers player decisions that could not be assigned to any defined decision. The confirmability of the feedback shall be supported by a graphic comparison of defined decisions with the decisions entered by the player.

T_4 **Challenge defined decisions:** The training module shall allow players to report decisions if they disagree with any of the expert-defined decisions presented during feedback.

T_5 **Personal performance dashboard:** The training module shall allow players to display a dashboard that presents their aggregated performance data, including decision accuracy, reaction time, and training intensity metrics. The dashboard shall also display video scenes with the lowest judgment performance regarding decision accuracy and reaction time.

T_6 **User performance comparison:** The training module shall allow players to view a leaderboard, their rank among other participants in the same course, and the average performance data of all players in a course.

T_7 **Mobile app support:** The training module shall allow players to access the prototype via a mobile app running on Android operating systems and optimised for 10-inch tablets.

T_8 **Decision Data Persistence:** The training module shall record all decisions entered by the player along with their determined correctness and reaction time metrics.

T_9 **Decision Matching:** The training module shall correlate the player's decisions with the decisions defined by the expert referee as a prerequisite to determine their correctness and reaction time.

Cross-Cutting Concerns Requirements

X_1 **Authentication methods:** The system shall allow users to log in via email/password, Google, or Moodle. Not yet registered users logging in via Google or Moodle shall only be allowed to access the training module.

X_2 **Moodle integration:** The system shall be able to be registered and started as an LTI (Learning Tool Integration) tool in the learning management system Moodle. For authorisation purposes, the already authenticated Moodle user shall be used.

X_3 **Extendability for other sports:** The system shall be easily expandable to other sports and disciplines. A configuration file shall be provided that allows flexible configuration of the disciplines' characteristics.

4.3 Exploratory Prototypes - Design Mockups

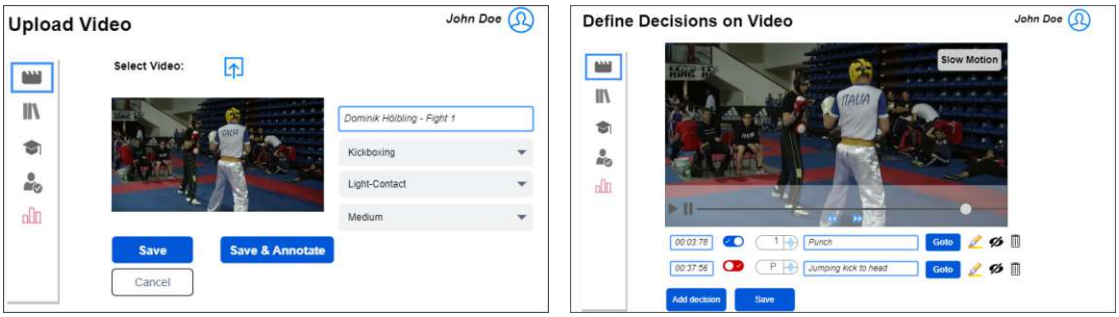
The mockups described in this section resulted from two iterations of exploratory prototyping, in which feedback from subject matter experts has been incorporated accordingly. These mockups aimed to envision the functionality of the content and administration module and the training module. Because they were designed based on an incomplete set of initially gathered requirements, they do not accurately reflect the design of the final prototype. Each of the subsequently presented mockups is described with reference to the requirements it covers. Changed or newly introduced requirements that resulted from the discussion of the individual mockups or intermediate high-fidelity prototypes from later iterations are also described.

4.3.1 Mockups for Content & Administration Module

This section presents the mockups created for the content and administration module, which covers features that allow authorized users to prepare and organize video scenes for serious game users and query selected performance statistics.

Upload Video and Define Decisions

The mockup shown in Figure 4.2a shows a first design draft for uploading a video along with basic metadata describing the characteristics of the presented fight (requirement C_1). Based on the uploaded video, users can annotate the video with decisions according to the events occurring in the fight, as shown in Figure 4.2b (requirement C_3). This mockup illustrates the feature of watching the video in slow motion, using a frame-by-frame function to precisely navigate the video and adding decisions described by the parameters of time, athlete, value, and technique. In addition, the functionality to highlight information-rich areas (requirement C_4) and to cover revealing referee gestures (requirement C_5) is indicated by icon buttons on the added decisions.



(a) Mockup: Video upload.

(b) Mockup: Definition of decisions.

Figure 4.2: Mockup: Video upload and definition of decisions.

Aside from the need to refine the fields describing the videos, discussions based on these mockups led to the realization that publicly available broadcast videos typically depict an entire bout that is too long to be used in the serious game. This resulted in the additional requirement (C_2) to cut a video into multiple video scenes before defining decisions on it. It turned out that the function to hide referees is only needed for point-stop disciplines, as referees in running-time disciplines do not indicate revealing gestures during the fight. In order to track the quality of the video scenes defined by subject matter experts, the requirement to introduce a simple status management was added in a later iteration (requirement C_6).

Compile Video Scenes in Playlists

The mockup shown in Figure 4.3 outlines a first variant of compiling video scenes in containers called "playlists". Playlists can be created by specifying a name and selecting items from the list of available video scenes (requirement C_7).

Discussions surrounding the functionality presented in this mockup suggested that more information needed to be included in the selectable video scene elements to distinguish them from one another. Due to the expected high number of available video scenes, a filter was requested that allows creators of playlists to limit the number of video scenes

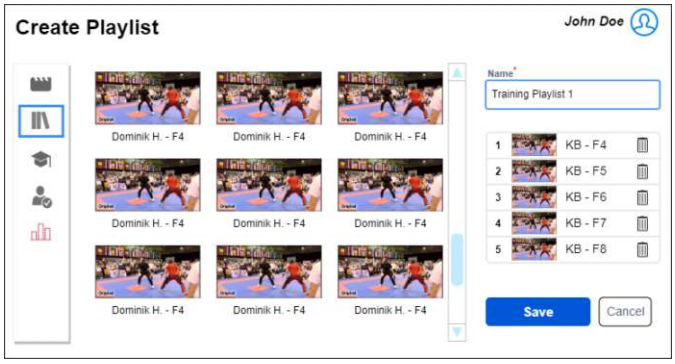


Figure 4.3: Mockup: Definition of a playlist including selected video scenes.

according to various criteria. In order to configure playlists for different use cases such as practice and scientific evaluations, a function for defining the repeatability of the playlist, the playback order of the video scenes it contains, and the extent of the displayed feedback was requested (requirement C_8).

Share Selected Playlist With Players

The mockup shown in Figure 4.4 sketches a first design variant for releasing selected playlists to players of the serious game in the form of a course (requirement C_9). In addition to its playlists and participants, a course is identified by its name, start date, end date, and the sports and disciplines covered. A course is available to participants for the period defined by the start and end date.

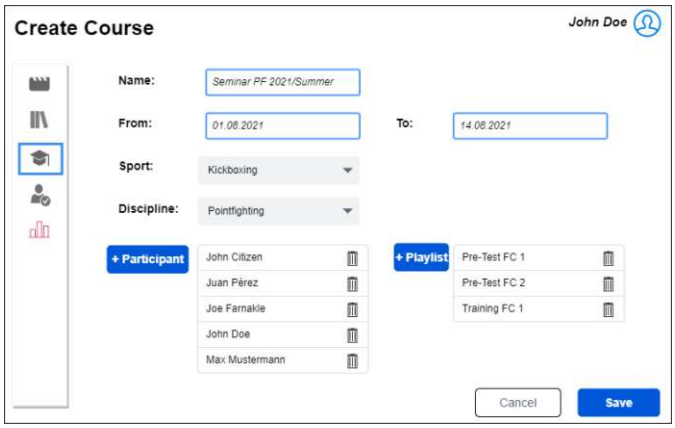


Figure 4.4: Mockup: Definition of a course including participants and playlists.

Discussions based on the outcome of later iterations showed the need to make certain video scenes available to all users of the serious game. This led to the introduction of a new requirement that allows publishing playlists for all serious game players (C_{10}).

Dashboard and Statistics

The mockup shown in Figure 4.5a outlines an early version of a dashboard that allows administrative users to monitor overall statistics on the performance of referees in selected courses (requirement C_{11}). It includes the users' overall decision accuracy and reaction time, as well as a ranking and the enumeration of video scenes with the lowest judgment performance in terms of decision accuracy and reaction time. For more detailed information on referees' performance, administrative users can query enhanced statistics by specifying sport, discipline, metric, and aggregation level (requirement C_{12}). Figure 4.5b exemplifies these features by showing the screen for querying the users' *decision accuracy* aggregated by *technique* for the discipline *Full contact* in the sport of *kickboxing*.

(a) Mockup: Administrator dashboard showing referees' performance data.

(b) Mockup: Statistics section to selectively query for the referees' performance.

Figure 4.5: Mockup: Administrator dashboard and statistics.

Discussions based on the functionality of this prototype suggested that the additional metric of training intensity should be provided on the dashboard. In addition, the performance data metrics should be aggregated by discipline. On the other hand, the display of a ranking was not considered relevant. Generally, the widgets displayed on dashboards should be role-specific to provide administrative users with relevant information at first glance.

Manage Users and Permissions

The mockup shown in Figure 4.6 outlines a draft of the screen that allows administrators to create users in the system. In addition to basic personal data describing the users, they are assigned roles defining which functions they are allowed to perform (requirement C_{13}).

Discussions and reflections based on this prototype slightly changed the structure of a user in the system. Instead of allowing multiple roles to be assigned per user, assigning just one role to each user was deemed sufficient. In addition, it should be possible for a

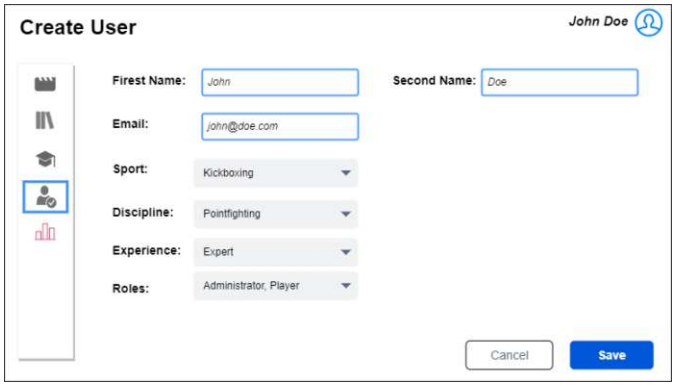


Figure 4.6: Mockup: Create user.

user to have several disciplines. The discipline and experience fields are only relevant to serious game players and not to administrative users.

4.3.2 Mockups for Training Module

This section introduces the mockups designed for the training module, covering features that allow referees to train their decision-making skills by following a video-based training approach.

Judge Video Scene

The mockup shown in Figure 4.7a outlines an early draft of the screen that presents a video scene for which the player is asked to attribute decisions to the blue or the red athlete using a simplified scoreboard (requirement T_1). Figure 4.7b shows another variant of this mockup, created for the mobile version of the serious game, with the scoreboard buttons on the left and right side of the video (requirement T_7). After judging the video scene, the player-entered decisions are persisted (requirement T_8) and correlated (requirement T_9) to the defined decisions as a prerequisite to compute performance data for feedback and statistical purposes.

Immediate Feedback

The mockup shown in Figure 4.8a shows a draft of the screen that provides feedback on the correctness of the judgments made by the player of the serious game (requirement T_2). The feedback includes the player's decision, the correct decision, the reaction time, the correctness of the decision, and the applied technique. In order to increase the confirmability of the decision, a 30 % slow-motion video sequence is presented, covering the period of 0.5 seconds before and after the moment of the defined decision. Players who disagree with the provided feedback can report it by entering a comment (requirement T_4). After presenting the feedback on each individual decision, a holistic feedback summary

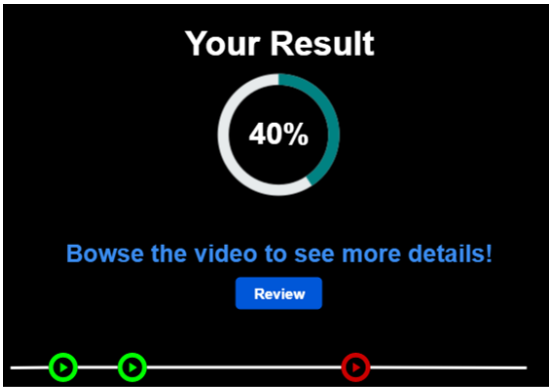
(a) Mockup: Running video scene to be judged by the player.

(b) Mockup: Running video scene to be judged by the player (mobile version).

Figure 4.7: Mockup: Video scene to be judged by the serious game player.

(see Figure 4.8b) is displayed that takes into account all of the player's decisions in the judged video scene (requirement T_3).

(a) Mockup: Feedback for individual decision.



(b) Mockup: Feedback for all decisions.

Figure 4.8: Mockup: Feedback on the players decision(s).

Personal Performance Dashboard

The mockup shown in Figure 4.9 outlines a first draft of the personal dashboard, including overall statistics on the logged-in referee's performance of selected courses (requirement T_5). It includes the user's overall decision accuracy and reaction time, as well as a ranking of the top three referees of the selected course (requirement T_6). In addition, it shows a listing of video scenes with the lowest judgment performance in terms of decision accuracy and reaction time.

Similar to the feedback on the administrator dashboard, discussions surrounding this functionality suggested that the additional training intensity metric should be provided in the dashboard, and the metrics should be aggregated by discipline. However, the

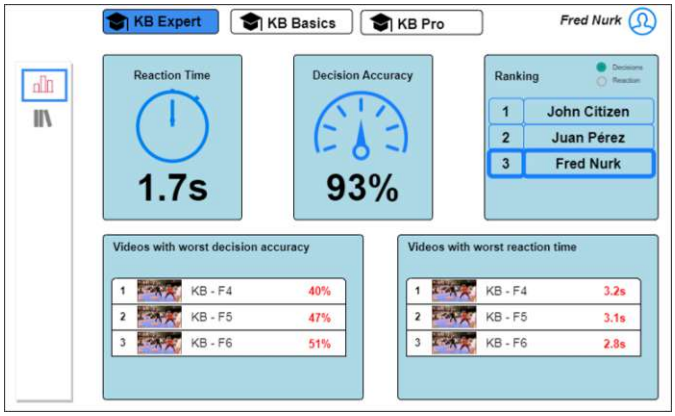


Figure 4.9: Mockup: Serious game dashboard showing personal performance data.

display of a ranking was refused. Instead, only the personal rank should be displayed along with a leaderboard presenting the average performance data of the best-performing referee within the selected course.

4.4 Functional Description of JudgED

This section provides a functional description of the developed serious game *JudgED* encompassing a training module as well as a content and administration module. While the training module is supposed to be used by referees to train their decision-making skills, the content and administration module allows authorised referees to define and organise the training videos used in the serious game. While both modules are provided as a web application, the training module is additionally accessible by an Android app. Each feature is described by referring to the requirements they implement.

4.4.1 Content & Administration Module

The content and administration module comprises a range of functionalities to (i) define the video-based training content used in the serious game, (ii) organise targeted training sessions accessible for certain players, and (iii) evaluate the players' performance data. This includes functions to upload videos, define video scenes, compose playlists, organise courses, and create performance statistics described in the subsequent sections.

Video Upload

Implemented Requirements: C_1

A video depicts a fight involving two athletes in a particular discipline. Videos can be uploaded along with basic metadata. Figure 4.10 shows the video creation screen consisting of a video file upload area (1) and descriptive metadata fields (2). While most

of the metadata fields exclusively serve descriptive purposes, the fields *association* and *discipline* determine the characteristics of the extractable video scenes.

The association defines the sport of the video and can take the values *WAKO Kickboxing* or *WKF Karate*. Admissible values for the remaining fields partly depend on the selected association. The discipline allows further specifying the sport depicted in the video. While the disciplines Point fighting, Light contact, Kick light, Full contact, Low kick, and K1 Style are selectable for the association WAKO Kickboxing, the association WKF Karate only supports the discipline Kumite. The purpose of the remaining fields bout, tournament, year, age group, athlete 1, athlete 2, and gender is to describe the context of the fight. The persistence of these fields allows for precisely identifying the video and contributes to high data quality by enabling the detection and prevention of duplicate videos.

Saving the video triggers the upload to the content delivery network *Vimeo*[1] and the persistence of the respective metadata in the JudgED database. As the duration of the upload process depends on the size of the selected video file, the video upload is performed in the background, allowing the user to continue working with the system while the upload is in progress. After the successful upload of the video, a transcoding process is initiated by Vimeo for optimisation purposes[2]. Figure 4.11 depicts the video overview page showing an uploading video (1), an uploaded video in the transcoding process (2), and already transcoded videos (3). The creation of video scenes on the uploaded video (4) is enabled as soon as the transcoding process is completed.

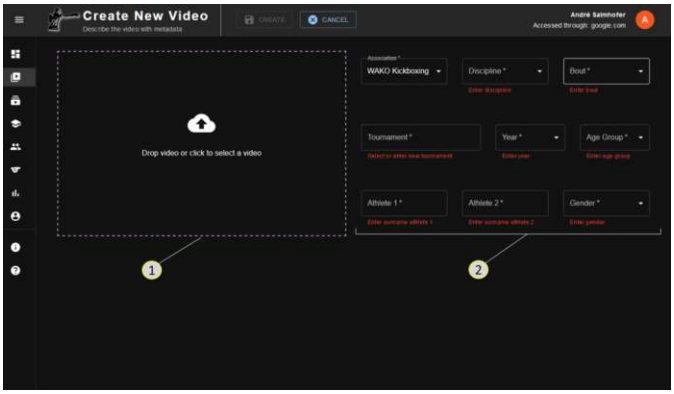


Figure 4.10: Upload of video including descriptive metadata.

Video Scene Definition

Implemented Requirements: C_2, C_3, C_4, C_5, C_6

[1] https://vimeo.com/
[2] https://developer.vimeo.com/api/guides/videos/upload

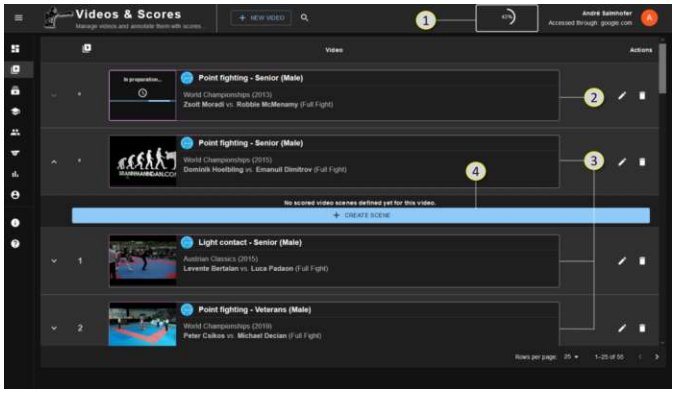


Figure 4.11: Video overview showing uploading, transcoding and already transcoded videos.

Once a video is uploaded to the system, and the transcoding process has finished, multiple video scenes can be extracted. A video scene is a segment of a video representing a fight situation that can be judged by the players of the serious game. As video scenes are annotated with a list of decisions appearing in the defined time range, they serve as a reference to (i) determine the correctness of the players' judgments, (ii) display appropriate feedback to the players, and (iii) calculate the players' performance data in terms of decision accuracy and reaction time. The subsequent paragraphs provide an overview of admissible configurations of video scenes and describe how they can be defined in JudgED.

Constraints: The constraints applicable to the duration and decisions of a video scene are determined by the discipline inherited from the uploaded video. *Point Stop* disciplines must include one decision and an optional concurrent decision within a duration of 4 to 15 seconds. In contrast, video scenes of *Running Time* disciplines require a duration between 45 and 90 seconds and can comprise multiple decisions. A decision can either include a numerical score value or a penalty. For simplification reasons, the penalties *warning* and *exit* (W and E) are consolidated. A special case is represented by decisions defined with a zero score. These decisions are supposed to mark sensitive situations for which no referee input is expected. Decisions must not be defined within the first second of the video scene, which allows players of the serious game to orient themselves in the fight scene. Prohibiting the definition of decisions within the last second and automatically appending a trailing period of two seconds at the end of each video scene ensures that players of the serious game have enough time to make decisions. Table 4.4 summarises the allowed configurations of video scenes with respect to duration, number of decisions, and the range of admissible decision values by sport and discipline.

Basic Configuration: Figure 4.12 illustrates the screen where video scenes can be defined. First, a fragment of the video is selected using the controls (1) to define the start and end time in the context of the uploaded video. The process of defining a

Table 4.4: Allowed configurations of video scenes in terms of duration, number of decisions, and the range of admissible decision values.

Sport	Discipline	Duration	Decisions	# Decisions
Kickboxing	Point fighting	4-15 s	0-3, W/E	1, 2
	Light contact	45-90 s	0-3, W/E	1+
	Kick light	45-90 s	0-3, W/E	1+
	Full contact	45-90 s	0, 1, W	1+
	Low kick	45-90 s	0, 1, W	1+
	K1 Style	45-90 s	0, 1, W	1+
Karate	Kumite	4-15 s	0-3, C1, C2	1, 2

decision is initiated by watching the selected video sequence and manually stopping the video at a point close to the occurring event. To allow the detection of more details while scanning the video scene, the playback speed can be toggled between normal and 30 % slow-motion (2). To precisely seek the exact moment of the occurring event, the frame-by-frame function allows forwarding and rewinding of the video in intervals of 0.02 seconds (3). Adding a new decision (4) automatically adopts the point in time of the currently visible video frame. Each decision includes the exact point in time (5), the athlete (red or blue) to which the decision is attributed (6), the technique leading to the respective score or penalty (7), and the decision value (8) automatically implied by the chosen technique. The position of the red and blue athletes must be configured (9) according to their position at the beginning of the video scene. This setting allows consistently assigning decisions to the particular athlete and properly aligns the red and blue scoreboard shown to the player in the serious game. To increase the confirmability of the decision, video scenes can be annotated with multiple tags describing the main focus of the included decisions, such as penalties, hits, punches or frequently occurring techniques. In addition, each video scene can be annotated with an overall judgement difficulty indicated by a five-point scale covering values from *very low* to *very high*.

Figure 4.12: Definition of video scene and occurring decisions.

Highlighting: Besides the basic configuration of video scenes described above, information-rich areas significant for detecting the correct decision can be highlighted for each defined decision separately. Figure 4.13 shows the drawing screen for a selected decision, where vital areas can be marked. A highlighting consists of one or more ellipses drawn on the video scene maximum 0.5 seconds before and after the defined decision (1). Within this time frame of one second, the visibility period of the highlighting can be adjusted appropriately to the situation's specifics (2). However, the visibility period must overlap with the point in time of the defined decision. Each ellipse can be positioned and resized to highlight certain techniques or crucial areas (3). The defined highlighting is displayed as an overlay in the decision-specific slow-motion feedback presented to the player in the serious game after judging the video scene, which aims to increase the confirmability of the decision.

Figure 4.13: Definition of highlighting for a defined decision.

Preview: The preview function allows reviewing each defined decision, along with the defined highlighting. It renders a 30 % slow-motion video sequence comprising the period 0.5 seconds before and after the defined decision, corresponding to the decision-specific feedback presented to the player in the serious game.

Blurring: Utilising videos from real-life competitions poses a problem, as referees indicating the potentially correct decision might appear in the video. While this problem does not exist for *Running Time* disciplines, where the fight is silently judged in the background, referees in *Point Stop* disciplines immediately stop the fight after each score or penalty by indicating an appropriate gesture. To avoid influencing the players in the serious game, referees can be covered by adding multiple blurring rectangles (1) for the time of the revealing gesture (2) as illustrated in Figure 4.14. Each blurring rectangle can be positioned and resized to ensure the referee's gesture is obscured (3). While the blurring rectangles are rendered transparently in the configuration screen to simplify the definition, they appear opaque for the serious game player.

Status Management: The content and administration module provides a simple status management to track the quality of video scenes. A newly created video scene is initially

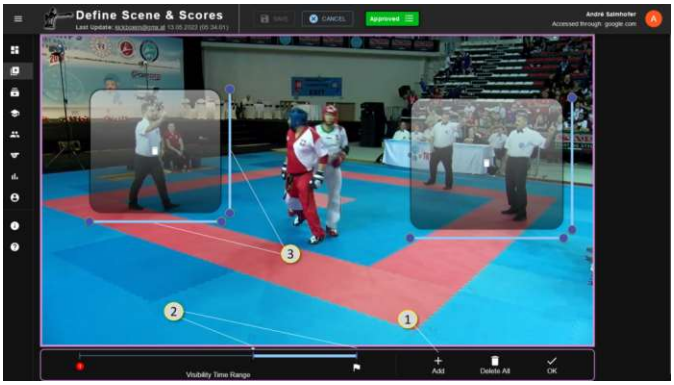


Figure 4.14: Covering decision-revealing gestures of referees.

in status *DRAFT* until reviewed by another authorised user, who can change the status to either *APPROVED* or *REJECTED*. Rejections require the user to enter the reason to allow the creator of the video scene to adapt it accordingly. Modifying decision-relevant fields of a video scene automatically triggers a status change to *DRAFT*. To enable the traceability of transitions, each status change is recorded with the user, date, and comment. Figure 4.15 shows the current status (1), the functions to approve (2) or reject (3) the video scene, as well as the history of status changes (4).

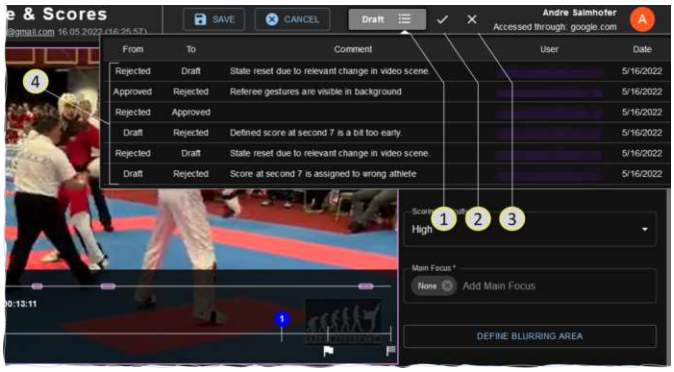


Figure 4.15: Status management of a video scene.

Playlist Creation

Implemented Requirements: C_7, C_8

A playlist is a container to compile a series of video scenes according to didactic or organisational requirements. The video scenes included in a playlist are supposed to be judged by the serious game player within one training session. As illustrated in Figure 4.16, playlists are composed by dragging and dropping (1) selected video scenes from left to the right. To estimate the compiled playlist's duration, the sum of the included video scenes' durations is indicated (2). Besides its included video scenes, a

playlist is characterised by its name as well as its mode, which can take the values *regular* (3), *lab* (4), or *exam* (5). While the name exclusively serves for identification purposes, the mode determines the playlist's behaviour with respect to allowed repetitions, playback order, and the extent of displayed feedback.

Figure 4.16: Playlist creation by drag and drop of video scenes.

Regular Playlists: These playlists are intended to be used for regular training sessions in non-scientific settings. Regular playlists can be played multiple times, whereas the included video scenes appear in random order. Comprehensive feedback is shown to the player after each video scene, and the slow-motion replay can be watched repeatedly.

Lab Playlists: These playlists are intended to be used for intervention periods in scientific settings. Like regular playlists, lab playlists can be played multiple times, their video scenes appear in random order, and comprehensive feedback is shown to the player after each judged video scene. However, compared to regular playlists, the slow-motion feedback is non-repetitive.

Exam Playlists: These playlists are used to examine the players' skills in scientific settings, such as pre-, post-, and retention tests in field experiments. Video scenes in exam playlists appear in their defined order and can only be judged once. No feedback is shown to the player after judging the video scene.

Course Creation

Implemented Requirements: C_9, C_{10}

Courses act as organisational units to temporarily unlock specific playlists for certain players of the serious game. Figure 4.17 shows the screen to create a course by defining the course participants (2), the accessible playlists (3), as well as the start and end date of the course (1). The configured course participants can access all playlists included in the course for the specified period. In addition, administrators can configure a course to be publicly accessible to all users with the disciplines specified in the course (4). If this option is selected, specific participants cannot be selected.

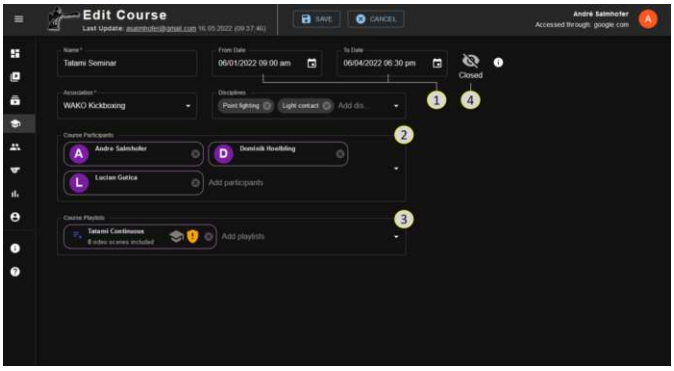


Figure 4.17: Course definition including playlists and players.

Administrator Dashboard

Implemented Requirements: C_{11}

The widgets visible on the dashboard are customised according to the role of the administrative user. While administrators can see statistics and charts comprising all players in the system, the data visible to course organisers are restricted to administered courses. Figure 4.18 shows the dashboard of a user with role *administrator* displaying the average decision accuracy (1), reaction time (2), and training intensity (3) in general and for each discipline separately. To analyse the trend of these metrics, a line chart provides insights about their development over time (4). To quickly detect video scenes posing problems for serious game players, the worst-rated video scenes in terms of decision accuracy and reaction time are enumerated (5). To indicate the quality of the video content, the number of video scenes challenged by players of the serious game (6) and the number of video scenes in status *REJECTED* (7) are displayed.

Figure 4.18: Administrator dashboard showing performance data.

Performance Statistics

Implemented Requirements: C_{12}

To precisely evaluate players' performance, the content and administration module allows for generating charts visualising decision accuracy, reaction time, and training intensity on multiple aggregation levels, such as user, discipline, or course. Figure 4.19 exemplifies the functionality by showing a chart for the metric reaction time (1) aggregated by discipline (2). Statistics can be created as bar charts (3) or line charts (4), visualising the trend of the respective metric over time. To refine the statistics, the data used for generating the charts can be filtered by various criteria (5).

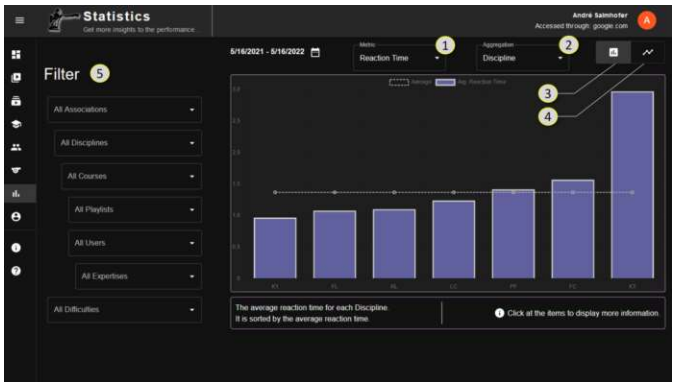


Figure 4.19: Chart showing the reaction time by discipline.

User Management and Permissions

Implemented Requirements: C_{13}

Administrative users can create and manage users within the association for which they are authorised. Figure 4.20 shows the screen for creating a user. The first section of the screen (1) contains basic personal information, the user's role, and the user's associations (scopes). The remaining fields (2) are only available for users of the role "player". They describe the player's disciplines and expert level for all associations selected above. Users created this way will receive an email containing a link they must follow to confirm their identity and set their initial password. After that, the initialisation process is complete, and the user will be redirected to the login page, where he/she can sign in to the application by entering the chosen password.

The authorisation concept of JudgED is based on the model of role-based access control (RBAC) [SCFY96] with the limitation that a user can only be assigned to one role. While the *role* determines the set of functions the user is allowed to access, the *scopes* define to which associations' data the functions are restricted. Figure 4.21 visualises the relationships between the supported roles player, administrator, organiser, and content creator regarding granted permissions. While players are only allowed to access the training module, other roles can access the content and administration module of the serious game. Administrators can create, read, update, and delete objects within their assigned scopes as long as data consistency is not violated. Organisers can create videos,

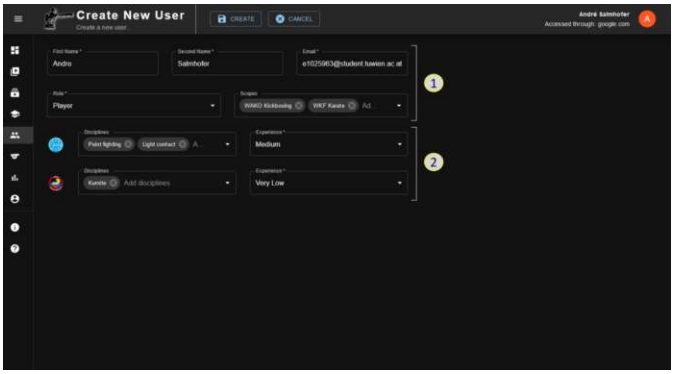


Figure 4.20: Creation of a user with role "player".

video scenes, playlists, and courses within their assigned scopes. However, they can only update and delete objects they have created themselves. Content creators have all the permissions of organisers but cannot create courses.

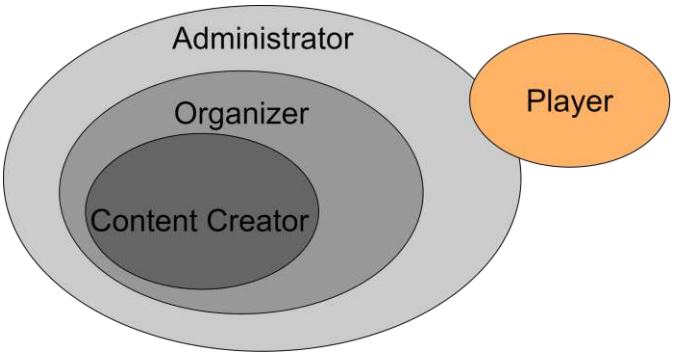


Figure 4.21: Relations between the roles player, administrator, organiser, and content creator.

4.4.2 Training Module

The training module provides a digital video-based training method to practice intuitive decision-making skills of martial arts referees by allowing them to judge numerous fights and get immediate feedback. The defined courses, playlists, and video scenes prepared in the administration and content module provide the basis to present the training sessions to serious game players.

Besides the web-based version of the training module, it is additionally provided as an Android mobile app optimised for ten-inch tablets. By allowing the players to indicate their decision on a touchscreen, the time between recognising the occurring event and actual user input is kept short. The subsequent sections describe the training module by exploring its mechanics and core functionalities.

Serious Game Mechanics

In the training module, the user takes on the role of the martial arts referee, who is confronted with a series of fight situations in the form of short video scenes. Using a discipline-specific scoreboard, the user's task is to make accurate judgments of occurring events as soon as they appear in the video to increase personal statistics. After each video scene, the serious game reveals the correct decision and renders a slow-motion video sequence highlighting information-rich areas leading to the score or penalty. To increase the users' motivation to train with the serious game, competitive elements such as personal statistics, rankings, and comparisons with other players are integrated.

This process intends to train the categorisation of perceived information through the method of multiple-cue probability learning [LNKS06]. By exposing referees to a large number of video scenes and providing immediate feedback, the aim is to establish relations between relevant cues (information perceived) and outcomes (the proper judgment for the situation at hand).

Judge Video Scene

Implemented Requirements: T_1, T_8, T_9

The user of the serious game initiates training sessions by selecting an available playlist, which starts one of the included video scenes. Figure 4.22 shows the running video scene, for which the player needs to judge appearing events in real time. Using a discipline-specific scoreboard, the player can attribute decisions to the blue or red athlete (1 and 2), increasing the current scores visible above the video (3 and 4). As the scoring rules are determined by the discipline and the age group of the fight, this information is displayed as well (5). While the footage progresses, potentially revealing referee gestures are hidden according to the blurring configuration of the video scene. To give players enough time to make decisions in case any assessable events occur near the end of the video scene, each video scene is extended by a trailing period of two seconds, indicated by a countdown.

Each judgement entered by the player is stored in the database, and its correctness and reaction time are calculated. Section 4.6 describes the data recording and performance measurement in detail.

Immediate Feedback

Implemented Requirements: T_2, T_3, T_4

After each video scene, the players receive immediate feedback on their decisions' correctness and reaction time as depicted in Figure 4.23. The extent of the displayed feedback is determined by the mode of the playlist in which the video scene is included (see section 4.4.1). For each decision defined in the video scene (1), the feedback comprises the player's decision (2), the correct decision (3), the reaction time (4), the correctness of the decision (5), and the applied technique causing the decision (6).

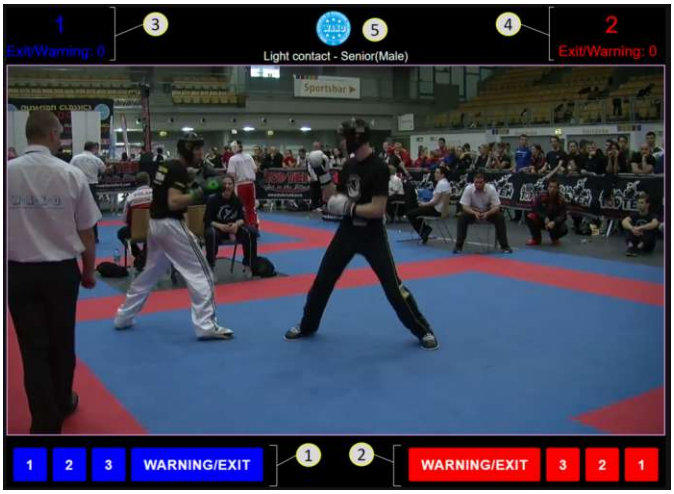


Figure 4.22: Video scene to be judged by the player.

To support the player's understanding of the provided feedback, a 30 % slow-motion video sequence comprising the period of 0.5 seconds before and after the point in time of the defined decision is shown to the player. To further increase the confirmability of the revealed decision, the slow-motion sequence optionally highlights important areas crucial for detecting the cause of the respective decision (7). If the player disagrees with any expert-defined decisions, they can be challenged by entering a comment explaining the reason for the disagreement (8).

The review closes with an overview summarising the performance of the judged video scene as shown in Figure 4.24. The summary includes the overall decision accuracy of the judged video scene and a visual comparison of the player's decision to the defined decisions on a timeline (1 and 2). Compared to the decision-specific feedback presented before, this summary also considers redundant decisions (3), which could not be matched to any defined decision.

Dashboard & Game Elements

Implemented Requirements: T_5, T_6

The dashboard depicted in Figure 4.25 serves as the landing page of the serious game. By visualising selected statistics indicating the personal judgment performance and stimulating the players' ambitions to improve on them, it aims to increase their motivation to train with the serious game intensively. The dashboard displays personal performance data in the form of average decision accuracy (1), reaction time (2), and training intensity (3) in general and for each discipline separately.

To compare the personal performance with the achievements of other players in the serious game, the dashboard displays a leaderboard (5), the player's rank (6), and the

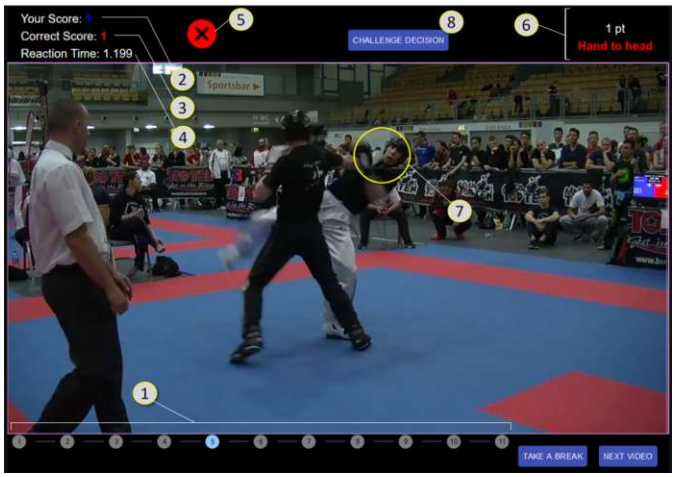


Figure 4.23: Feedback about the player's decision(s) after judging the video scene.

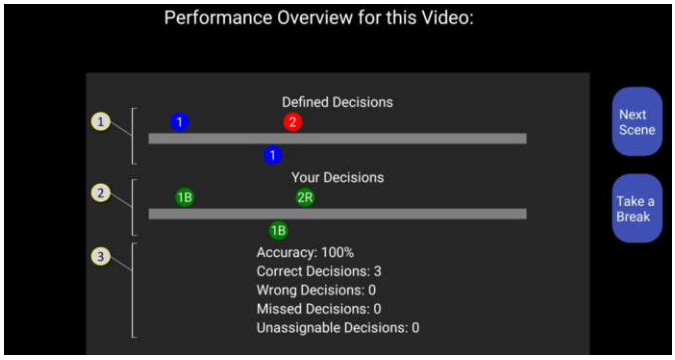


Figure 4.24: Holistic feedback about the judged video scene.

average performance data among all players (7). While the leaderboard indicates the decision accuracy and reaction time of the best player in the selected course, the own rank refers to the rank of the currently signed-in player. Both leader board and own rank are determined based on the metric of decision accuracy. To avoid frequently fluctuating data in these elements, players with less than 50 decisions are not considered, as the amount of recorded performance data is assumed to be too little for an expressive performance indication.

Besides the performance statistics described above, the dashboard also lists video scenes for which no decision was correctly assessed by the player (4). This widget lets players selectively replay video scenes and improve their statistics. They disappear from the list as soon as the player replays these video scenes and reaches a decision accuracy greater than zero percent.

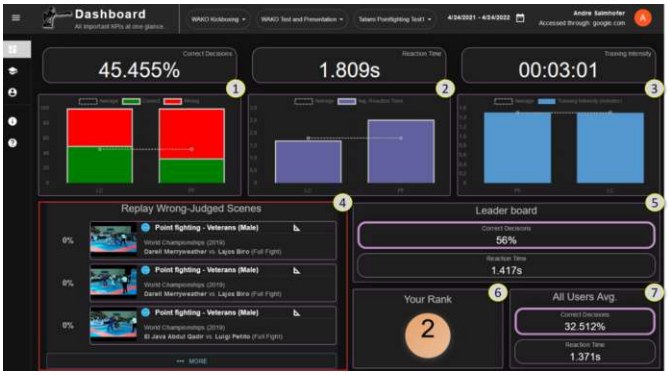


Figure 4.25: Game dashboard including personal performance data and comparison with other players.

Serious Game Mobile App Support

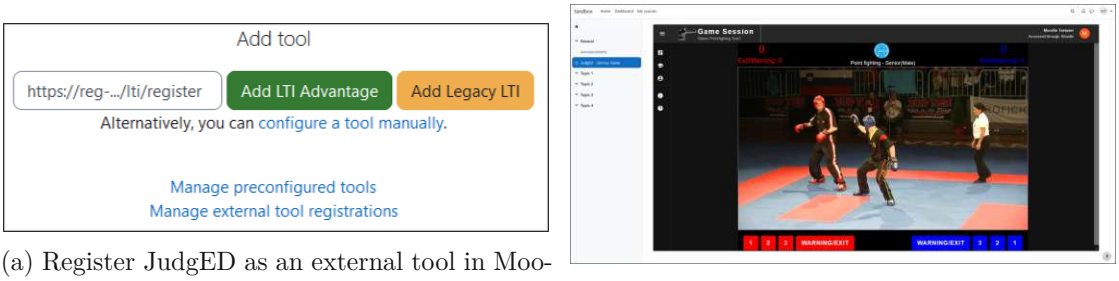
Implemented Requirements: T_7

The training module is also provided as an Android mobile app optimised for 10 inch tablets. While most mobile screens are similar to the web interface, the scoreboard buttons for judging the video scenes are placed on the left and right sides of the video, respectively, for better ergonomics. In addition, this screen is only supported in landscape mode to take advantage of the screen ratio for maximising the size of the rendered video. Figure 4.26 shows the judging screen as it appears in the mobile app version of the serious game.

Figure 4.26: Video scene to be judged by the player on a mobile device.

4.4.3 Cross-Functional Concerns

Compared to the sections above, which describe functions attributable to either the content and administration module or the training module, this section presents general functions relevant to both modules.



(a) Register JudgED as an external tool in Moodle.

(b) Starting JudgED via Moodle.

Figure 4.27: Registration and start of JudgED in Moodle.

Authentication and Integration

Implemented Requirements: X_1, X_2

The serious game allows users to log in using three different authentication methods: (i) email/password, (ii) Google, and (iii) Moodle. Users who are not yet registered and log in via Google or Moodle are automatically created with the "Player" role. Therefore, they can only access the training module by default. While the login through email/password and Google is initiated from the JudgED login page, launching JudgED via Moodle requires registering it as an external tool on the respective Moodle installation. Registration is done through LTI Advantage's dynamic registration process [ltia], which only requires providing the JudgeED registration URL as shown in Figure 4.27. After successful registration, JudgeED can be embedded in Moodle courses as an external tool and started with the currently logged-in Moodle user as depicted in Figure 4.27b.

Configuration of Sports and Disciplines

Implemented Requirements: X_3

In order to flexibly expand JudgeED with new sports, the structure and characteristics of associations and disciplines are defined by a configuration file in JSON format. The configuration file defines associations and their disciplines, as well as the properties of disciplines in terms of name, scoring scheme, allowed video scene duration, and applicable techniques. JudgeED uses this information to render discipline-specific scoreboards, customise selectable fields in drop-down menus, and perform plausibility checks. This way, new sports and disciplines can be supported by adapting the configuration file and minimal code changes.

4.5 Architecture & Applied Technologies

This section provides an overview of the data structure, architecture, and technologies used to develop and run JudgED. After describing the data structure, including its main entities and their relationships to each other, the architecture of the system and its

components that constitute the functionality of JudgED are presented. Subsequently, the infrastructure on which the prototype is deployed is described.

4.5.1 Data Structure

To define and deliver the video content for the serious game players, the prototype is based on the main entities of videos, video scenes, playlists, courses, and users. The training content of the serious game is based on uploaded *videos* depicting fight situations between two athletes. By including information about the associated video file stored in Vimeo, videos serve as a basis to render the footage to be judged by the player of the serious game. As many publicly available videos include an entire bout, a video is partitioned into multiple *video scenes* corresponding to fight sequences supposed to be judged by the player of the serious game. To arrange a series of video scenes according to didactic or organisational requirements, they are compiled in the form of *playlists*. To make selected playlists available for certain groups of referees, *courses* serve as organisational units combining playlists and *users*. Users assigned to a course can access all included playlists for a defined period. Figure 4.28 illustrates the relationship between the entities described above.

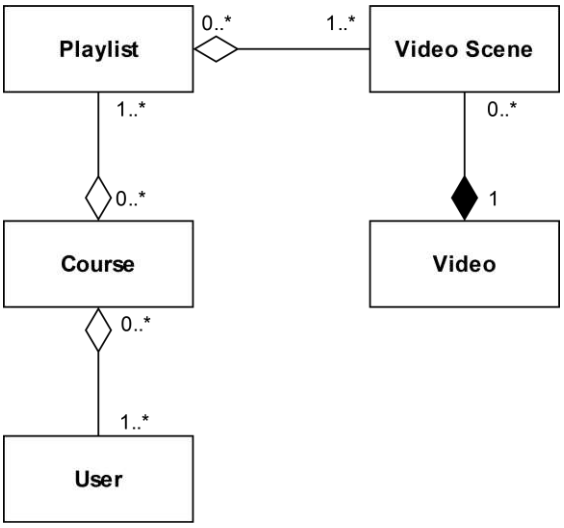


Figure 4.28: Relation between main entities of the serious game.

Instances of these entities are persisted in the form of JSON-like objects in the NoSQL database MongoDB. While the paragraph above provides a rough overview of the involved entities and their relationships, Appendix 7 provides more insights by describing the *JSON Schema* of each entity.

4.5.2 Architecture & Components

The system was designed by considering the requirements gathered during the requirements engineering process. From a bird's eye view, the system is structured into the layers of presentation, business, and storage. The presentation layer of the prototype was developed by applying component-based software engineering [XLKR00]. The business layer was developed following a resource-oriented architecture [Ove07] exposing data retrieval endpoints to be consumed by the frontend. The endpoints were implemented according to principles of API composition and API aggregation [BCC+17], which allowed to keep the client-side code slim by encapsulating complexity in the backend. The content delivery network Vimeo [vima] was utilised to store and stream the training videos. The system's architecture was realised by applying the MERN stack [Sub19], including the technologies MongoDB, Express, React, and NodeJS.

Figure 4.29 provides an overview of the components involved in establishing the functionality of JudgED. The diagram is organised into groups that contain coherent components. The subsequent sections describe these component groups by referring to their purpose, interactions, and the technologies used to implement them.

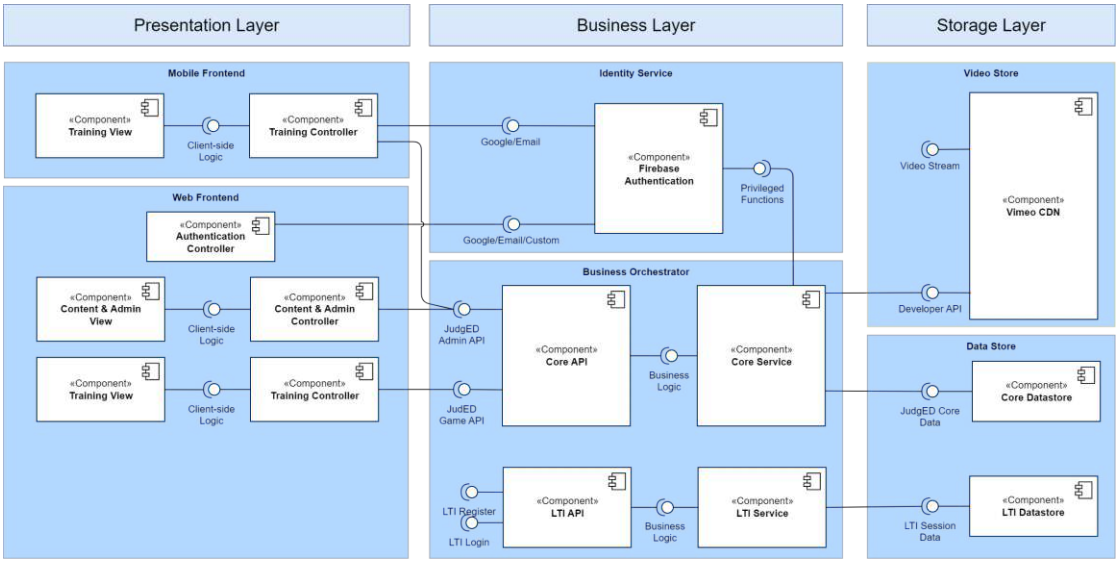


Figure 4.29: Overview about components involved in JudgED.

Web Frontend

The web frontend was developed as single page web application, which aims to bring the look and feel of native applications to the web browser by utilising JavaScript, HTML, and Cascading Style Sheets [Sco15, p. 3]. It communicates with the business layer to execute certain functions and retrieve data to be displayed. The web frontend was built using the open-source, component-based JavaScript library React [reaa]. By creating

complex components composed of multiple simple components managing their own state, React enabled the creation of advanced user interfaces. The creation of the interactive web application was supported by features such as state and lifecycle management, event handling, and conditional rendering of components [reaa].

The creation of the web frontend was supported by various third-party JavaScript libraries providing client-side routing functionalities, advanced graphical components, internationalisation capabilities, and other utility functions. The client-side routing library *React Router* [rou] allows mapping the application's main components to defined routes of the application, which causes the rendering of the associated components as a response to user requests. The graphical component library *MaterialUI* [mui] provides standard components such as buttons, text fields, dropdowns, and sliders, which were customised according to the needs of the individual use cases. Supported by the library *React Intl* [intb], all texts visible in the user interface are translatable into various languages. While an English translation is provided in this work, additional languages can be supported by adding separate JSON files, including the translations in the form of key-value pairs. The package manager NPM [npm] was used to install external JavaScript packages from an online repository. The file package.json includes the list of all packages used in the web application and their version.

The web frontend is logically structured into the content and administration (cf. 4.4.1) and training components (cf. 4.4.2). To separate the definition of the user interface from the client-side business logic, the frontend of the serious game and the content and administration module is divided into the layers view and controller. The responsibility of the view layer is to render the user interface, including its dynamic data, as requested by the user. The controller layer serves as an intermediary to facilitate communication between the view layer and the business layer. Functions of the controller layer trigger HTTP requests towards the business layer's REST endpoints and update the state with the returned data, which re-renders the view accordingly. While the view layer is built with React function components [reac], the controller layer is implemented in the form of React custom hooks [reab]. Before allowing users to access these REST endpoints, the authentication controller handles the sign-in process of the user by communicating with the identity service, which provides a token that can be used to verify the user's identity and access rights in the business layer.

Mobile Frontend

The frontend of the training module is additionally provided as an Android mobile app developed by the React-based JavaScript framework React Native [read]. Besides sharing the same programming language, React and React Native use the same mechanism for state management. Whenever the state of a component changes, its view gets automatically re-rendered [Eis15]. Due to these similarities and the separation of user interface and client-side business logic according to the same architectural principles, a part of the code developed for the React web frontend could be reused in the mobile

frontend. While the view needed to be implemented from scratch using React Native components, controllers developed for the web application were reused.

Business Orchestrator

By exposing REST endpoints that retrieve data from various sources and execute dedicated functions to fulfil certain use cases, the business orchestrator provides services to the frontend accessible for authorised users. It is implemented in JavaScript and executed in the open-source runtime environment Node.js [nod]. It was designed according to resource-oriented architecture principles [Ove07] supported by the web application framework Express.js [expa], which provides HTTP routing and middleware utilities for creating robust endpoints. Each endpoint is defined by its associated HTTP method, path, and callback function. The application listens for incoming HTTP requests and executes the callback function matching the respective HTTP method and path definition [expc]. In addition, the router-level middleware utility provided by Express.js allows injecting reusable functions in the request-response cycle [expb]. This is used to check the users' authorisation before executing the callback function.

The functionality of the business orchestrator is logically structured into core and LTI (Learning Tools Interoperability) components. While the core components are responsible for providing services to the web and mobile frontend, LTI components allow to register JudgED as an external tool in Moodle-based environments and start it by using the authenticated Moodle user. The implementation of the LTI components is supported by the Ltijs library [ltib], which offers utilities allowing JudgED to be registered and started as LTI 1.3 Learning Tool [ltia] in Moodle.

To separate the protocol-specific definition of the REST endpoints from the actual implementation of the business logic, both core and LTI components are separated into the layers of API and service. This allows reusing functions of the service layer throughout the application without being limited by protocol-specific characteristics. The API layer handles HTTP requests from the frontend towards defined REST endpoints and delegates the execution of the business logic to designated functions of the service layer, which communicate with other components to perform the requested tasks. If the execution succeeds, a positive response is returned to the frontend containing the result in JSON format. Otherwise, depending on the source of the error, the respective HTTP error code is returned to the frontend along with a descriptive error message.

Identity Service

The identity service is realised by Firebase Authentication [firb], which offers services to manage the identity of users. By providing functions for signing-in users, verifying their identity, and persisting them along with custom attributes, this component serves as a basis for authentication and authorisation in JudgED. While the app development platform Firebase [ird] provides a multitude of services, only Firebase Authentication is

used in this work, which allows users to authenticate via password, phone numbers, and common federated identity providers like Google or Twitter [firb]. Despite these standard authentication methods, the custom authentication method allows authenticating users with systems not directly supported by Firebase.

Authentication methods must be selectively enabled in Firebase before they can be used. The Firebase project used in this work allows users to authenticate via the methods of Google, password, and custom. While the Google method allows users to sign in with their existing Google account through OAuth 2.0 [oau], the password method requires the user to register for an account beforehand. The custom authentication method is used for automatically signing in the user in case JudgED was started as an external tool from a Moodle-based environment.

To develop these authentication flows, Firebase provides a client SDK [firc] offering functions that can be invoked from untrusted clients and an admin SDK [fira] that provides utilities for interacting with Firebase from privileged environments. When the user triggers the sign-in using one of the authentication methods provided by the frontend, the respective function of the client SDK is called to authenticate the user. If the authentication process succeeds, a signed JWT token containing basic user information is returned and stored in the client's session. Besides basic user information like user ID, name, and email, the token also contains the user's access rights as stored in the custom attributes of the associated Firebase user record. The token is provided as a parameter in all HTTP requests from the frontend towards the business layer. As the issued JWT token is signed with the private key of the Firebase project, its integrity can be verified in the business orchestrator, which allows securely retrieving the identity and access rights of the logged-in user. This way, the business orchestrator checks whether the user is authorised to execute the respective function for each HTTP request.

Authentication via Moodle is a special use case implemented using Firebase's custom authentication method. Figure 4.30 visualises the authentication flow used when launching JudgED as an external tool in Moodle. Moodle initiates the process by calling the `/login` endpoint exposed by the JudgED API. Moodle provides the ID and email address of the authenticated Moodle user as parameters for this request. In the next step, the JudgED endpoint requests the generation of a custom token from Firebase Authentication for the provided Moodle user ID. Then the user is redirected to the `/moodleSignIn` URL of the JudgeED frontend with the user's email address, name and custom token as parameters. The JudgeED frontend uses the custom token to log into Firebase and update the auto-generated Firebase user with the email and name parameters provided. In response, Firebase returns a valid JWT token. If this is the user's first login via Moodle, the role needs to be initialised in the Firebase user's custom claims, which are propagated to the JWT token once a token refresh is requested in the frontend. The authentication process is completed by redirecting the user to the JudgED landing page.

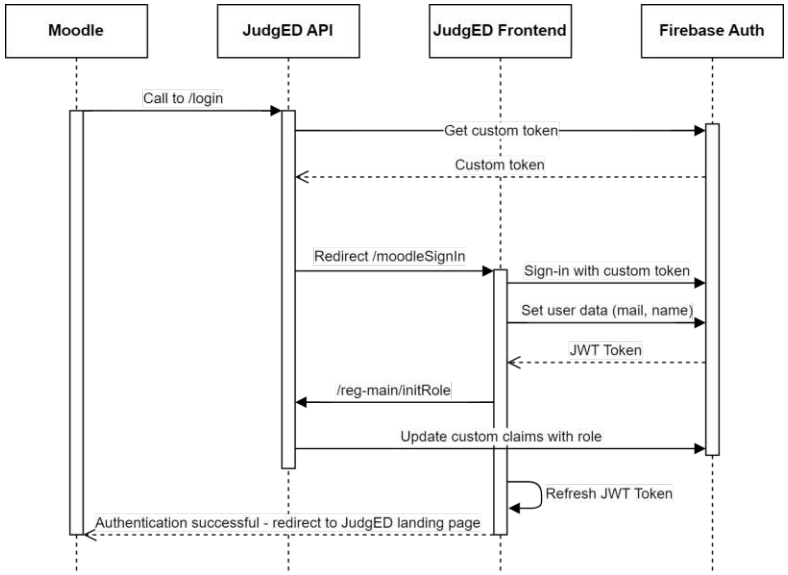


Figure 4.30: Using Firebase custom authentication to access JudgED from Moodle.

Data Store

Apart from the videos stored in the video store component, the data store holds all data objects required by JudgED. This component enables the business orchestrator to create, read, update, and delete data as requested by the user in the frontend. While the growth of most entities is estimated as low, the logging of the users' judgment performance is expected to accumulate large amounts of data over time. Therefore, the database must be able to persist and query large datasets efficiently. The application of document-based databases is a possible solution to fulfil these requirements, as they offer better scalability than relational databases in case of rapidly growing volumes of data [SAZ+18]. The data store is implemented with the document-based database MongoDB [monb].

The data store consists of two separate MongoDB databases: Core datastore and LTI datastore. The core datastore is responsible for persisting the objects required to define and deliver the video content for the serious game players. It is organised in multiple collections corresponding to the entities described in the section 4.5.1. While the schema and content of the core datastore are directly influenced by the self-written procedures of the business orchestrator's core service component, the LTI datastore is managed indirectly. It is exclusively used by the Ltijs library [ltib] in the LTI service component to persist session data in case JudED was started via the learning environment Moodle.

Video Store

The video store component is realised by the content delivery network Vimeo [vima], which stores and delivers all video files used in JudgED. To manage videos, the Vimeo developer API [vimd] exposes REST endpoints used by the business orchestrator to generate video

upload links, delete videos, create thumbnails, and request certain information about uploaded videos.

While most operations to manage video files are directly triggered by the business orchestrator towards the developer API as requested by the user in the frontend, the video upload process used by the content and administration module involves a two-step approach. First, to upload a video file to Vimeo, the business orchestrator requests a temporary upload link using the Vimeo developer API, which is returned to the frontend. Then, using this upload link, the actual upload is performed on the client side via the *tus* upload method [tus]. This reduces the time necessary for the upload process, as video files do not need to be passed through the business layer before being uploaded to Vimeo. After successful upload, the video gets moved to a dedicated folder within the Vimeo account to separate the videos according to the stage from which the video was uploaded (development, staging, production). In addition, the configurations are applied to the video in the form of Vimeo embed presets [emb]. At the end of this process, the link of the uploaded video is stored as an attribute in the respective video entity persisted in the data store component.

Videos are uploaded with the privacy settings "unlisted" and "public". While "unlisted" means that videos do not appear in any Vimeo search and can only be accessed with the link, the attribute "public" makes them accessible for everyone on the internet [vimb]. This way, the video stream can be directly accessed via the frontend using the video link persisted in the data store's video object.

4.5.3 Deployment Infrastructure

While the section 4.5.2 describes the structure of JudgED by enumerating its main components and applied technologies, this section provides an overview of the infrastructure on which the application is deployed. Figure 4.31 shows a deployment diagram consisting of nodes, artefacts, and deployment specifications required to establish the functionality of JudgED. The separation of development, staging, and production environment is illustrated by encoding the different staging environments in square brackets of the nodes. This notation was used to reduce redundancy in the diagram. Communication is only enabled between nodes of the same staging level. For naming the nodes and artefacts, the abbreviation "reg" (referee education game) is used, which was the initial project name before it was renamed to "JudgED". The following paragraphs describe the scope and structure of each main node depicted in the deployment diagram.

Client Devices

The training and content and administration modules are provided as single-page web applications accessible via the web browser on the client's workstation. The training module of JudgED is also accessible via an Android mobile device by installing the respective APK (Android Package) file. Depending on the applied build settings to generate the

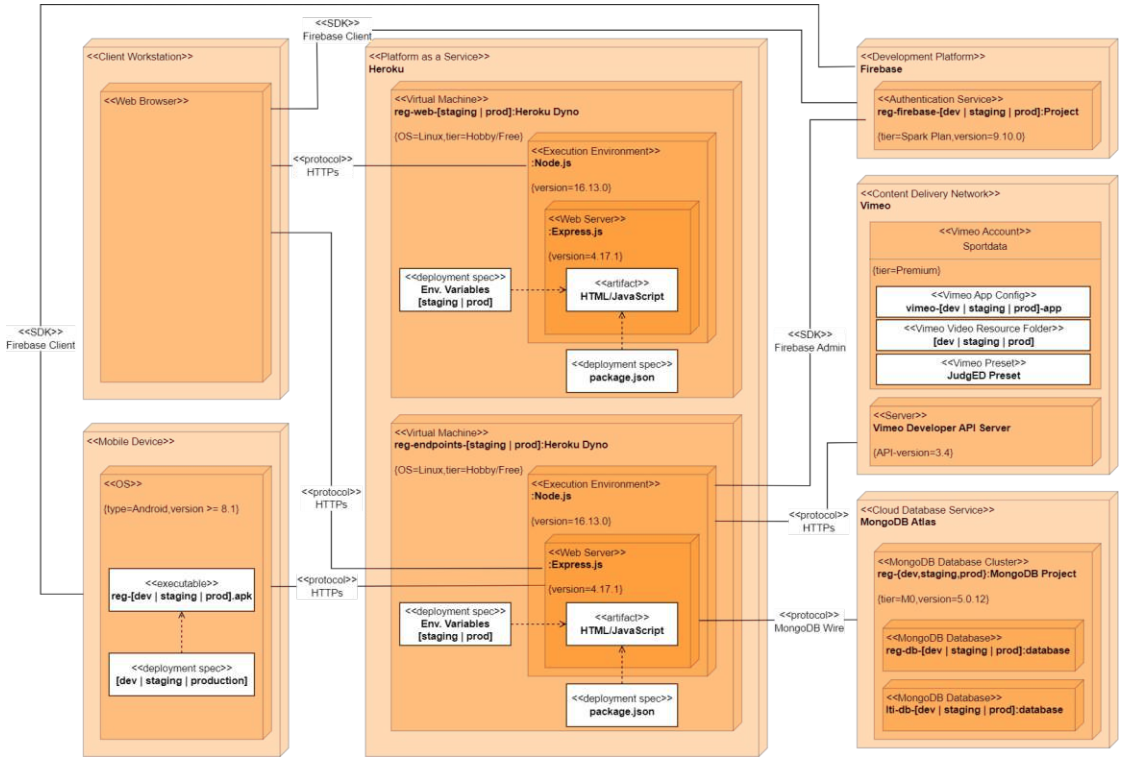


Figure 4.31: Deployment diagram describing the infrastructure required to run JudgED.

APK file and the URL used to request the web content, the client communicates with the development, staging, or production backend node to call the exposed REST endpoints via HTTPs. The production installation file of the mobile application was published in the Android Play Store, which eased the installation by not requiring users to edit their security settings on their mobile phones with respect to accepting third-party APK files.

Heroku Dynos

The static web content, as well as the business logic exposed as REST endpoints, is deployed on the Platform as a Service (PaaS) provider Heroku [hera]. In Heroku, the source code and dependencies are packaged and deployed into "dynos", which are isolated environments providing an operation system, memory, and computation power [herb]. The static web content is served by a dyno (reg-web-*) providing a Node.js runtime environment [nod], which utilises Express.js [expa] to handle requests triggered from the client's web browser. Using the same technology stack, another dyno (reg-endpoints-*) handles HTTP requests from the mobile application and web browser towards the exposed REST endpoints providing the application's business logic.

To distinguish between the staging and production environment, two dynos are used for serving the static web content (reg-web-staging, reg-web-prod) and two dynos for exposing

the REST endpoints (reg-endpoints-staging, reg-endpoints-prod). The environment variables configured on the respective dynos determine with which external system the application communicates. The environment variables in reg-web-* dynos determine the Firebase project and the reg-endpoints-* dynos with which the client communicates. The environment variables in reg-endpoints-* dynos determine the Firebase project, Vimeo account, and MongoDB database cluster to which the dyno connects.

The staging environment uses dynos of tier "Free", which become idle after 30 minutes of interactivity. Dynos of the production environment use the tier "Hobby", which eliminates this disadvantage by being always active [herb]. The deployment process to these dynos is automatised with the help of the DevOps platform Gitlab [git] used in this project. Pushing changes to the *staging* respectively *master* branch of the Git repositories copies the files to the respective Heroku dyno and deploys the application.

Firebase

To allow users to authenticate in JudgED via Firebase Authentication, separate Firebase projects of the free Spark tier [fire] were created for the development, staging, and production environment (reg-firebase-*). Among other attributes, each project generates a Firebase project ID, an API key, and a Firebase authentication URL, which allows untrusted nodes (i.e. client workstation, mobile device) to communicate with the respective Firebase project by using the Firebase client SDK [firc]. In addition, a service account including a private key is generated for each project, which allows privileged nodes (i.e. reg-endpoints-*) to communicate with Firebase via the Firebase Admin SDK [fira].

The projects are configured to allow authenticating users via the methods Google, Email/Password, and Custom. Instead of creating separate records for users logging in with the same email address using multiple authentication methods, the projects are configured to link accounts based on the email address used. This allows authenticating users with the same email address through multiple authentication methods. As Firebase Authentication also provides utilities to confirm email addresses and reset passwords, the configuration also comprises email templates.

Vimeo

To store and stream videos, the Vimeo account of the company Sportdata [spo] is used. As this account uses the Vimeo "Premium" plan, it has sufficient capacity to store the videos used in JudgED. To allow nodes to communicate with the Vimeo API, Vimeo requires registering the application (i.e. Vimeo app) from where the endpoints are supposed to be called. To distinguish between the development, staging, and production environments, separate Vimeo apps were registered. A unique access token exists for each Vimeo app, which needs to be provided in each request towards the Vimeo developer API. This enables Vimeo to authenticate API requests and correlate them to the registered app. Each access token is attributed with scopes determining the range of permitted actions

[vimd]. The access tokens used in this work were generated with the scopes private, edit, delete, upload, public, and video files [vimc].

Although creating separate Vimeo apps for each environment logically separates the API calls, it does not determine the storage folder and settings applied to uploaded videos. To better organise uploaded videos, separate folders for the environments development, staging, and production were created in the used Vimeo account. To define reusable configuration templates, Vimeo allows defining embed preset [emb], which can be applied to uploaded videos. An embed preset corresponds to a collection of settings defining the behaviour and appearance of the Vimeo video player in the user interface when embedding the video stream. As all functions to interact with the video player are controlled programmatically, embed presets are configured in a way that does not allow the user to control the video in the user interface actively. The folder and the embed preset are applied by the respective API call after the video is successfully uploaded.

MongoDB

The cloud database service MongoDB Atlas [monb] is used for hosting the MongoDB databases required by JudgED. To distinguish between development, staging, and production environments, different database tier M0 clusters [mond] were created, which provide a storage capacity of 500 MB. Each cluster contains the core database (reg-db-*) and the LTI database (lti-db-*). Clients can communicate with the databases through the TCP/IP-based MongoDB Wire protocol [mone]. Accessing the databases is only allowed if the respective user credentials are provided. Network communication with the cloud clusters is allowed from every source IP over the internet.

4.6 Data Recording & Performance Measurement

This section describes what data is logged in JudgeED to enable the analysis of player performance. After describing the procedure and quality of data recording in JudgED, the algorithm for determining the correctness and reaction time of the user's decisions is presented.

4.6.1 Data Recording

Each decision is logged in the database to analyse the players' performance and generate statistics. Particularly, the following attributes are persisted for each judged video scene as a basis to generate expressive statistics: (i) user ID, (ii) video scene ID, (iii) playlist ID, (iv) course ID, (v) date and time of the judgment, and (vi) a list of player decisions including input time, decision value and athlete.

The recording and persistence of these fields form the basis for calculating the players' performance data in terms of decision accuracy and reaction time. To obtain expressive results, the time of the player inputs must be precisely recorded in accordance with the

progress of the streaming video scene. This is enabled by the functions of the client-side video player library used to render and interact with the video stream. There are two basic methods to determine the current progress of the video stream: (i) by listening for `timeupdate` events triggered in an interval of 250 ms or (ii) by explicitly invoking the function `getCurrentTime()`.

Depending on the capabilities provided by the client-side video library used for the web and mobile interface of the serious game, either the first or the second approach is utilised. While the web version of the serious game actively invokes the *getCurrentTime()* function whenever the player makes a decision, the mobile version relies on asynchronously processing *timeupdate* events, which comes with a maximum inaccuracy of 250 ms. Independent of the used library, this allows the conclusion that the serious game can record the player's decisions in accordance with the progress of the streaming video by tolerating a maximum inaccuracy of 250 ms.

4.6.2 Performance Measurement

To display appropriate feedback at the end of video scenes and to statistically evaluate the players' performance data on various aggregation levels, the accuracy and reaction time must be determined for each individual decision. While decision accuracy is defined as the percentage of correct decisions, the reaction time of a decision indicates the elapsed time between the player's decision and the defined decision.

To determine these metrics, the entered player decisions are compared to the decisions defined in the respective video scene. In case multiple decisions are defined in the video scene, it is sometimes unambiguous which player inputs were intended for which defined decision. An algorithm was developed to solve this problem, which correlates player decisions to defined decisions based on a set of rules.

Definitions

Before scrutinising the correlation process in detail, the following paragraph defines basic terms used throughout the explanation of the algorithm.

Defined Decisions $\mathcal{D}$*:* A list of decisions included in the video scene specified by an expert referee in the administration and content module. Each defined decision $\mathcal{D}_i$ includes the properties *time*, *athlete* and *decision value*.

Player Decisions $\mathcal{P}$*:* A list of decisions entered by the serious game player while watching the video scene. Each player decision $\mathcal{P}_j$ includes the properties *time*, *athlete* and *value*.

Matching $\mathcal{M}$*:* A tuple representing the correlation of a player decision $\mathcal{P}_i$ with a defined decision $\mathcal{D}_j$. As additional information, each object contains the reaction time and the correctness of the decision.

Unassignable Decision $\mathcal{P}_u$*:* A subset of player decisions not assignable to any defined decision.

Missed Decision $\mathcal{D}_u$*:* A subset of defined decisions not assignable to any player decision.

Maximum Decision Time $\mathcal{T}_{max}$*:* The maximum admissible period between player decision and defined decision to be considered as a matching $\mathcal{M}$. It is defined as three seconds.

Considerations for Choosing $\mathcal{T}_{max}$

The maximum period $\mathcal{T}_{max}$ between player decision and defined decision is a necessary but insufficient condition for a decision to be considered correct. For the comparable video-based training platform SET [SPKB11] developed to train the decision-making skills of soccer referees, a time range of five seconds was used. While the referred study used a mouse as an input device, the present work recommends using the mobile version of the serious game, in which the players indicate their decisions by tapping on the respective scoreboard button on the touchscreen. Using the touch screen, the time between recognising the occurring event and entering the decision is assumed to require less time than using a mouse input device. Therefore, the threshold of three seconds for $\mathcal{T}_{max}$ is considered appropriate for the serious game in this work.

Correctness Evaluation

Player decisions are only considered correct in case (i) they are judged within $\mathcal{T}_{max}$ seconds from the point in time of the defined decision, (ii) they are entered in the same order as the defined decisions, and (iii) the athlete and decision value correspond to the defined decision.

To enable the evaluation of decisions by considering the criteria mentioned above, the serious game uses an algorithm which correlates player decisions to defined decisions according to a defined set of rules. This algorithm results in a list of matching tuples $\mathcal{M}$ fulfilling the conditions (i) and (ii). Condition (iii) can be checked by comparing the athlete and decision value of the player decision with the defined decision for each matching in $\mathcal{M}$. The following section describes the approach for generating the list of matchings as a first step to determine the correctness of decisions.

Decision Matching Procedure

1. Basic matching: The list of matched decisions $\mathcal{M}$ is produced by correlating player decisions in $\mathcal{P}$ with defined decisions in $\mathcal{D}$. Player decisions are attempted to be matched with defined decisions fulfilling the correctness and maximum decision time condition. If no matching candidate satisfies the correctness condition, the defined decision is assigned to the nearest player decision. Already correlated player decisions are not allowed to be used for further correlations.

Defined decisions with value *zero* pose a special case, as they require no explicit player input to be considered as correct. If the defined decision is a zero-score and no player inputs are assignable, it is assigned to a synthesised player decision that is constructed by taking the time, athlete, and decision value properties of the defined decision. The correctness of the generated matching is set to true. As the time parameter of the synthesised player input has no meaning, the reaction time property is not calculated for these matchings.

2. Conflicts Detection: Performing the procedure described above results in a list of matched decisions fulfilling the maximum decision time constraint. However, the correct order condition of player decisions might be violated after the basic matching step. Matchings causing the violation of this condition are referred to as *conflict*. A conflict represents a constellation in $\mathcal{M}$, in which a player decision's time parameter is smaller than the player decision's time parameter corresponding to one of its preceding defined decisions.

By chronologically arranging defined decisions and player decisions on two parallel timelines and graphically indicating their matching, a conflict can be detected by intersecting links. Figure 4.32 illustrates an example of a conflict involving the matching $\mathcal{M}_{1,2}$ and $\mathcal{M}_{2,1}$, where $\mathcal{M}_{ij}$ represents the matching of a defined decision at index i in $\mathcal{D}$ with a player decision at index j in $\mathcal{P}$. Formally, two matchings $\mathcal{M}_{ab}$ and $\mathcal{M}_{xy}$ are causing a conflict, if the conditions Equation 4.1 to Equation 4.5 are satisfied.

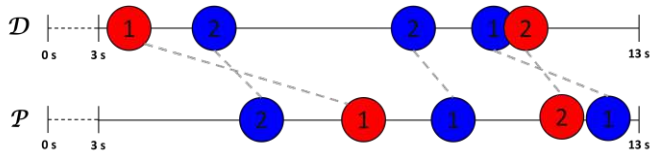


Figure 4.32: Conflict indicated by the intersection of matchings.

The consideration of Equation 4.5 slightly relaxes the definition of a conflict for decisions defined within a period of 0.3 seconds referred to as *delta time* δ. An incorrect order of player decisions does not cause a conflict for a sequence of decisions defined within delta time. Demanding the exact order for judging consecutive events appearing in such a short period was considered too strict, as it does not reflect constraints applicable in real-life contests. The intersection depicted by the matchings $\mathcal{M}_{4,5}$ and $\mathcal{M}_{5,4}$ illustrated in Figure 4.32 do not cause a conflict, assuming that $\mathcal{D}_4$ and $\mathcal{D}_5$ are defined within delta time.

$$\mathcal{P}_b.time - \mathcal{D}_y.time < \mathcal{T}_{max} \tag{4.1}$$
$$\mathcal{P}_y.time - \mathcal{D}_a.time < \mathcal{T}_{max} \tag{4.2}$$
$$\mathcal{D}_a.time < \mathcal{D}_x.time \tag{4.3}$$
$$\mathcal{P}_b.time > \mathcal{D}_y.time \tag{4.4}$$
$$\mathcal{D}_x.time - \mathcal{D}_a.time > \delta \tag{4.5}$$

3. Conflict resolution: To ensure the player decisions fulfil the correct order condition, conflicts identified in the previous step need to be resolved. A conflict is resolved by examining two matches involved in the conflict and determining which one to keep and which to refuse. Apart from eliminating the refused matching, its defined decision is added to the missed decisions $\mathcal{D}_u$, and its player decision is added to the unassignable decisions $\mathcal{P}_u$. This conflict elimination process is iteratively applied to all conflicts until the matching list $\mathcal{M}$ is free of conflicts.

The rule to decide which matching to keep is determined by the *obviousness* of the defined decisions involved in the conflict. A defined decision with a higher value is considered more prominent and therefore more likely to be correctly recognised. If the values of the defined decisions are equal, the matching with the earlier defined decision is kept. Equation 4.6 and Equation 4.7 define the degree of obviousness for decision values in kickboxing and karate Kumite, respectively. While the numbers in the expressions represent score values, *C1* (category 1), *C2* (category 2), and *Warning/Exit* correspond to penalty categories.

$$3 > Warning/Exit > 2 > 1 \tag{4.6}$$
$$3 > C2 > C1 > 2 > 1 \tag{4.7}$$

4. Outcome: The primary result of the matching procedure is the list of matched decisions $\mathcal{M}$ fulfilling the maximum time and order maintenance constraint. As a side product, the algorithm produces the lists of missed decisions $\mathcal{D}_u$ and unassignable decisions $\mathcal{P}_u$. Applying this procedure to the player decisions and defined decision depicted in Figure 4.32 would produce the following lists:

- $\mathcal{M}$**:** ($\mathcal{P}_1$,$\mathcal{D}_2$), ($\mathcal{P}_3$,$\mathcal{D}_3$), ($\mathcal{P}_4$,$\mathcal{D}_5$), ($\mathcal{P}_5$,$\mathcal{D}_4$)
- $\mathcal{D}_u$**:** $\mathcal{D}_1$
- $\mathcal{P}_u$**:** $\mathcal{P}_2$

Decisions in $\mathcal{D}_u$ and $\mathcal{P}_u$ are inherently incorrect and have no reaction time. In contrast, elements in $\mathcal{M}$ are enriched with information about their correctness and reaction time

determined by comparing their linked decisions. While the correctness of a matching is determined by comparing the athlete and decision values, the reaction time is calculated by building the difference of the decisions' time parameters.

CHAPTER 5

Evaluation

This chapter reports the outcomes of the evaluation of JudgED based on the data collected during the field experiment and the responses from the survey. It presents the results of the field experiment conducted during the Austrian WAKO Kickboxing Championship 2022 in Graz, which tested the decision accuracy and reaction time of licensed kickboxing referees in the serious game. The results obtained from statistical analysis of the recorded data from the field experiment are discussed in section 5.1. The subsequent survey collected additional information about the referees who participated in the field experiment. The findings obtained from statistical analysis of the responses are discussed in section 5.2.

5.1 Analysis of Field Experiment

This section discusses the results obtained from the descriptive and inferential statistical analysis of the recorded data from the field experiment. After reporting the overall performance data in terms of decision accuracy and reaction time, the probability of deriving a correct decision within a referee team and the level of agreement among the referees on their judgments is presented. Subsequently, the differences between the two tests performed during the field experiment reported and the results obtained from the four-choice reaction time test are analysed for correlations with the reaction time in JudgED.

5.1.1 Average Decision Accuracy and Reaction Time

The results of the field experiment showed a mean decision accuracy of 43.011 % (minimum = 27.047 %, maximum = 61.517 %, $\sigma = \pm$ 12.898 %) and a mean reaction time of 1.022 s (minimum = 0.755 s, maximum = 1.299 s, $\sigma = \pm$ 0.156 s) considering all referees participating in both tests. A Shapiro-Wilk test showed the normal distribution of the

users' decision accuracy, W(16) = 0.889, p = 0.053, $\alpha = 0.05$ and reaction time, W(16) = 0.942, p = 0.372, $\alpha = 0.05$. Figure 5.1 visualises the decision accuracy grouped by discipline, which accents the higher decision accuracy of Tatami disciplines (50.460 %) compared to Ring disciplines (30.596 %).

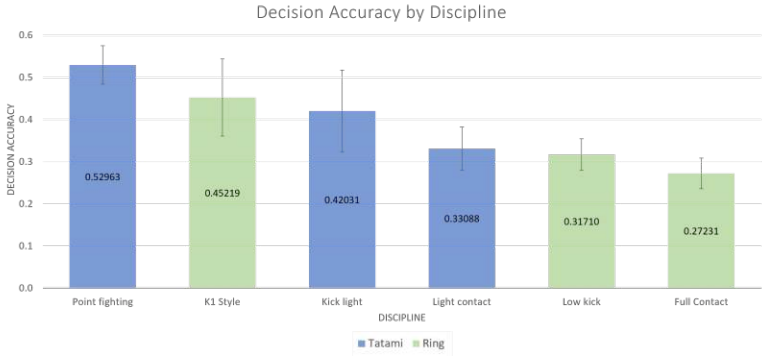


Figure 5.1: Decision accuracy by discipline for Tatami (blue) and Ring (green).

The data set was also analysed for the decision value of the defined decision as depicted in Figure 5.2. 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 %.

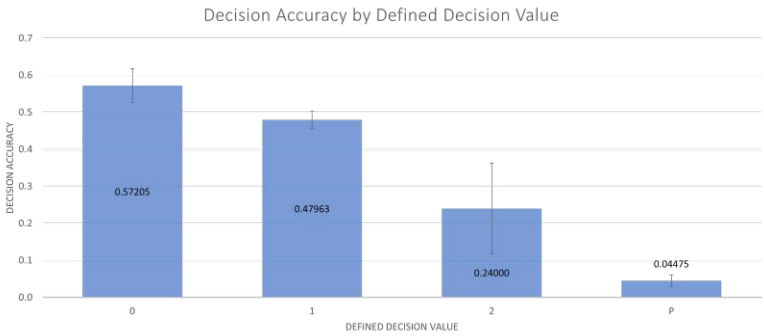


Figure 5.2: Decision accuracy by defined decision value.

5.1.2 Probability of Correct Majority Decision

Considering that certain decisions in professional WAKO kickboxing competitions require a majority decision among three independent judges [waka], the probability that at least two of them judge a particular event correctly was calculated using a binomial distribution formula. Let X be the number of referees making the correct judgment for a particular event, then the $P(X \geq 2) = P(X = 2) + P(X = 3)$. Based on the average

Table 5.1: Overall results from the Fleiss' kappa analysis for Ring referees.

	Kappa	Asymptotic Std. Error	Z	P	Lower 95 % Asymptotic CI Bound	Upper 95 % Asymptotic CI Bound
Overall	.371	.009	41.617	.000	.353	.388

Table 5.2: Individual results from the Fleiss' kappa analysis for Ring referees.

Rating Category	Cond. Prob.	Kappa	Asymp. Standard Error	Z	P	Lower 95 % Asymp. CI Bound	Upper 95 % Asymp. CI Bound
0-blue	.660	.644	.015	42.986	.000	.615	.673
0-red	.679	.663	.015	44.239	.000	.633	.692
1-blue	.458	.315	.015	21.031	.000	.286	.344
1-red	.480	.325	.015	21.666	.000	.295	.354
No input	.650	.362	.015	24.129	.000	.332	.391
P-blue	.031	.024	.015	1.578	.114	-.006	.053
P-red	.050	.041	.015	2.763	.006	.012	.071

decision accuracy of each discipline, the probability of at least two out of three judges making a correct individual decision is as follows: Point fighting (54.439 %), K1 Style (42.850 %), Kick light (38.148 %), Light contact (25.599 %), Low kick (23.789 %), Full contact (18.207 %).

5.1.3 Inter-Rater Reliability

The Fleiss' kappa, κ, was calculated to measure the level 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 effective subjects (defined decisions) from 6 raters (referees). The kappa value of 0.371 indicates a poor strength of agreement between the referees that is statistically significantly different from zero ($p < 0.0005$). Table 5.1 lists the results from the Fleiss' kappa analysis for Ring. Further data analysis reveals the differences in agreement between individual rating categories as shown in Table 5.2. While the kappa coefficients for the decision value *zero* (0.644, 0.663) and *one* (0.315, 0.325) indicate a moderate respectively poor level of agreement, decisions with value *penalty* (0.024, 0.041) show an agreement that lies only slightly above chance agreement. The rating category *no input* refers to defined decisions with values other than zero for which referees did not indicate a judgement.

Table 5.3: Overall results from the Fleiss' kappa analysis for Tatami referees.

	Kappa	Asymptotic Std. Error	Z	P	Lower 95 % Asymptotic CI Bound	Upper 95 % Asymptotic CI Bound
Overall	.398	.008	50.377	.000	.383	.414

Table 5.4: Individual results from the Fleiss' kappa analysis for Tatami referees.

Rating Category	Cond. Prob.	Kappa	Asymp. Standard Error	Z	P	Lower 95 % Asymp. CI Bound	Upper 95 % Asymp. CI Bound
0-blue	.735	.726	.014	53.132	.000	.699	.753
0-red	.690	.674	.014	49.356	.000	.648	.701
1-blue	.491	.326	.014	23.876	.000	.299	.353
1-red	.525	.332	.014	24.304	.000	.305	.359
2-blue	.185	.173	.014	12.636	.000	.146	.199
2-red	.242	.235	.014	17.223	.000	.209	.262
No input	.644	.452	.014	33.091	.000	.425	.479
P-blue	.167	.161	.014	11.784	.000	.134	.188
P-red	.000	-.005	.014	-.371	.711	-.032	.022

Tatami: The Fleiss' kappa test for Tatami was based on the judgment of 119 effective subjects (defined decisions) from 10 raters (referees). The kappa value of 0.398 indicates a poor strength of agreement between the referees that is statistically significantly different from zero ($p < 0.0005$). Table 5.3 lists the results from the Fleiss' kappa analysis for Tatami. Further data analysis reveals the differences in agreement between individual rating categories as shown in Table 5.4. While the kappa coefficients for the decision value *zero* (0.726, 0.674) indicate a moderate level of agreement, decisions with value *one* (0.326, 0.332) and *two* (0.173, 0.235) indicate a poor level of agreement. Decisions judged as *penalty* (0.161, -0.005) show an agreement that lies only around chance agreement. The rating category *no input* refers to defined decisions with values other than zero for which referees did not indicate a judgement.

5.1.4 Differences between Test 1 and Test 2

Scrutinising the decision accuracy and reaction between both tests showed that the results worsened from test 1 to test 2. While decision accuracy decreased from 45.333 % (± 17.194 %) to 40.910 % (± 10.149 %), reaction time increased from 1.008 s (± 0.168 s) to 1.036 s (± 0.176 s). Further analysis showed that the reduction in the decision accuracy was mainly caused by participants of the Tatami disciplines. While the majority of Ring referees (83.3 %) increased their decision accuracy, the majority of Tatami referees (80

%) worsened their results from test 1 to test 2.

The Pearson correlation coefficient was computed to test for the linear correlation between test 1 and test 2. The result indicates a significant strong positive correlation between the participants in test 1 and test 2 for decision accuracy, $r(14) = 0.798$, $p < 0.001$ and reaction time, $r(14) = 0.658$, $p = 0.006$. This shows that participants ranked high in test 1 were also ranked high in test 2.

To examine the differences between test 1 and test 2 of the field experiment, a dependent t-test was computed. To test the assumption for performing the t-test, a Shapiro-Wilk test was computed. The differences of the paired values were approximately normally distributed for the measure of decision accuracy, $W(16) = 0.949$, $p = 0.469$, $\alpha = 0.05$ and reaction time, $W(16) = 0.975$, $p = 0.910$, $\alpha = 0.05$. The result of the t-test indicated no significant differences between both tests for decision accuracy, $t(15) = 1.614$, $p = 0.127$, $\alpha = 0.05$ and reaction time, $t(15) = -0.776$, $p = 0.450$, $\alpha = 0.05$.

5.1.5 Correlation with Difficulty and Refereeing Experience

Further analysis of the performance data indicates a correlation between the difficulty rank of the 664 judged video scenes with decision accuracy and reaction time. Figure 5.3 visualises the referees' performance by the difficulty of video scenes. To show the strength and significance of the correlation, Spearman's ρ was calculated on the judged video scenes, which indicates a weak negative relationship between difficulty rank and decision accuracy, $r(662) = -0.281$, $p < 0.001$ and a very weak positive relationship between difficulty rank and reaction time, $r(563) = 0.99$, $p < 0.019$. For 99 of the video scenes, no average reaction time was available, as no player decision was assignable to defined decisions. Therefore, the latter result was only based on 565 items.

The Pearson correlation coefficient was computed to test for the linear correlation between the performance of referees and their refereeing experience in years. The results indicate no significant relationship between refereeing experience and decision accuracy, $r(12) = -0.388$, $p = 0.171$. Likewise, the results show no significant relationship between refereeing experience and reaction time, $r(12) = -0.196$, $p = 0.503$.

5.1.6 JudgED Reaction time and Four Choice Reaction Time Test

The mean reaction time resulting from the four-choice reaction time test was 0.545 s (± 0.074 s). To examine the relationship between participants' reaction time in JudgED and the four-choice reaction time test, the Pearson correlation coefficient was computed, which showed no significant linear relationship between the results, $r(14) = -0.132$, $p = 0.627$.

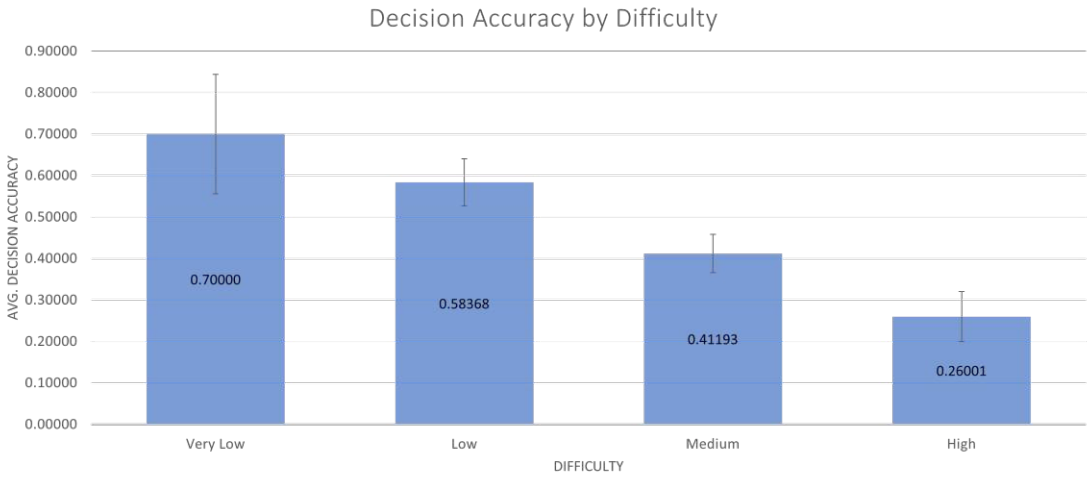


Figure 5.3: Decision accuracy by difficulty.

5.2 Analysis of Questionnaire

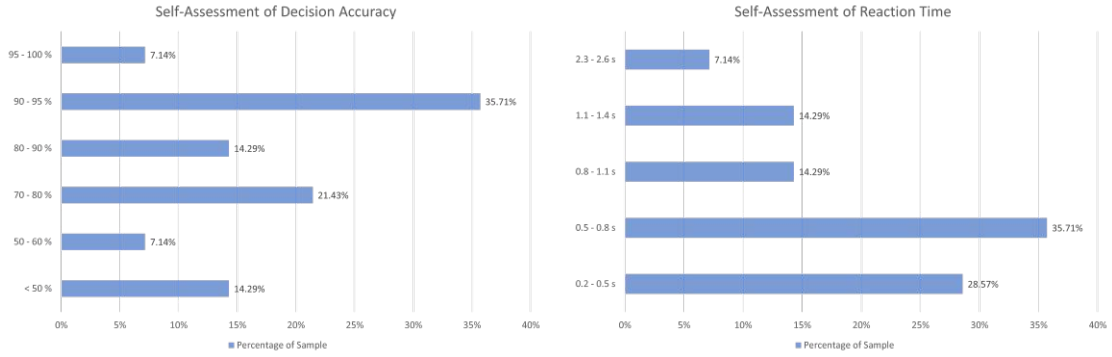
This section presents the results of the survey conducted after the field experiment using the questionnaire described in Appendix 7. The analysis of the questionnaire covers the topics (i) self-estimation of refereeing performance, (ii) experience with computer-supported referee education, (iii) attitudes towards current and future functions in JudgED, and (iv) potential, target groups, and application areas of JudgED. Subsequently, the general feedback provided by the respondents is discussed.

5.2.1 Self-Estimation of Referees

The referees who took part in the survey rated their general refereeing performance as very good (21.43 %), good (57.14 %) and average (21.43 %). Referees state they are confident in their refereeing skills (85.71 %) and their ability to make accurate judgments (92.86 %). They feel up to the challenge of meeting the demands placed on them (100 %). The high level of self-confidence is underpinned by the high self-assessment of decision accuracy and reaction time, as shown in Figure 5.4a and Figure 5.4b. The Spearman's ρ was computed to test a possible relationship between the referees' self-assessed performance rank and performance in the field experiment. The results show no significant correlation with decision accuracy, $r(12) = -0.108$, $p = 0.713$ and reaction time, $r(12) = 0.217$, $p = 0.457$.

5.2.2 Computer-Supported Referee Education

The results from the questionnaire showed that only 35,71 % of the referees participating in the survey had already used a computer program during their referee education. The used applications comprised e-learning platforms to share slides, rule explanation



(a) Self-assessment of decision accuracy. (b) Self-assessment of reaction time.

Figure 5.4: Self-assessment of referees' performance in competitions.

software, software to analyze fight situations, and software to complete theoretical referee exams. Despite the low exposure to computer programs during referee education, 85,71 % believe that the use of digital tools can improve referee education in general and has the potential to contribute to better learning outcomes. While 57.14 % of respondents say that computer programs are useful in improving their knowledge of rules, 71.43 % believe that computer programs can help improve their practical judgments.

The introduction of competitive elements in referee training was viewed negatively by the referees. Only a minority of referees indicate that anonymous comparison of acquired refereeing skills among colleagues would increase learning motivation (21.43 %) and enjoyment (28.57 %) during referee education. Likewise, only a small proportion of referees state that the introduction of personalized rankings can increase learning motivation (28.57 %) and enjoyment (21.43 %) during referee training.

5.2.3 Statistics and Features in JudgED

The judges unanimously emphasized the importance of making quick decisions when judging the fight. With an agreement of 85.71 %, they also show how important it is not only to judge fight situations in isolation but also to take into account previous observations during the fight. The referees confirmed the usefulness of displaying statistics in JudgeED regarding reaction time and decision accuracy with an agreement of 78.57 % and 57.14 %, while only 7.14 % denied their usefulness. The usefulness of comparing one's statistics with statistics from other referees was evaluated neutrally. While 28.57 % agree or disagree, 42.86 % neither agree nor disagree.

A section of the questionnaire was dedicated to collecting opinions on implementing certain features related to the scope of rankings, the playback of slow-motion feedback, the duration of video scenes, and the layout of decision buttons. Like the referees' rejection concerning the introduction of competitive elements in referee training, the usefulness of displaying ranking lists in JudgeED was also doubted. While referees would

limit the scope of the rankings to referees within a course (42.86 %) or referees of the same disciplines (21.43 %), the majority of referees (50 %) indicated that rankings are not necessary at all. The characteristics of the slow-motion sequence, shown after the video scene was assessed, were rated as reasonable in terms of speed (71.43 % agreement) and duration (78.57 % agreement). The appropriateness of the duration of the video scenes in *Point Stop* and *Running Time* disciplines was confirmed with an agreement of 71.43 % and 57.14 %, respectively. However, a non-negligible proportion of referees (21.43 %) consider *Point Stop* videos to be too short and *Running Time* videos to be too long. The referees were ambivalent about the arrangement of the buttons for assessing warnings and exits. While 50 % believe a common button is sufficient, 42.86 % believe separate buttons should exist.

5.2.4 Potential, Target Groups and Application Areas of JudgED

Referees believe that using JudgED during referee education can enhance classes (78.57 %) and improve learning outcomes (71.43 %). In particular, 57.14 % of respondents agree that JudgED can enhance the practice-oriented part of on-site referee training, while 21.43 % believe that JudgED has the potential to replace it. Furthermore, the referees believe training with JudgeED is helpful for both inexperienced (78.57 %) and experienced judges (71.43 %) to expand their skills. Additionally, 57.14 % agreed on the usefulness of JudgED as a tool to improve kickboxing athletes' knowledge of the rules.

As visualized in Figure 5.5, referees believe that the future application area of JudgED is (i) during courses for practical exercises, (ii) after courses as a complementary training method, and (iii) independently of courses, in order to keep skills up-to-date. Most referees would use JudgED on a tablet device (64.29 %). Referees' responses to the voluntary use of JudgED indicated a low willingness to train regularly with JudgED (7.14 % every few days, 14.29 % once a week, 42.86 % less than once a week, 35.71 % not at all).

5.2.5 Open Feedback

Optional feedback from an open question at the end of the questionnaire provided suggestions for improving the experience and future versions of JudgED. While one referee pointed out the importance of having a sufficiently large screen to detect athletes' movements, another mentioned that video scenes used in JudgED should be filmed from different perspectives. One referee mentioned that the position of the red and blue athlete must be more prominently highlighted at the beginning of the video scenes.

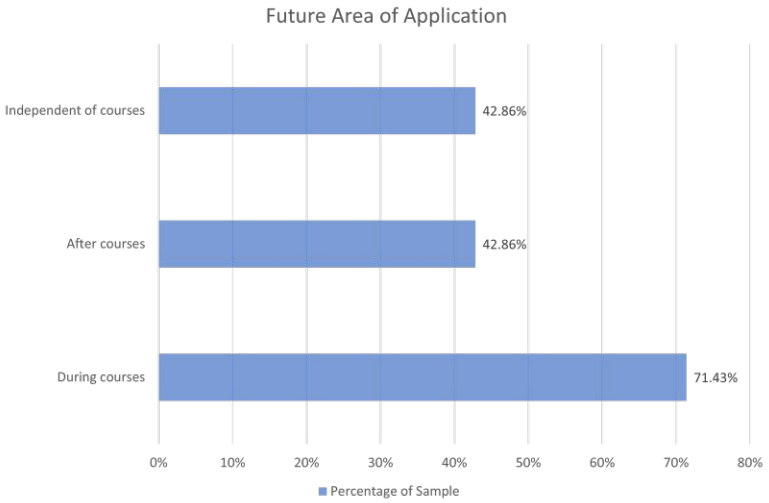


Figure 5.5: Future application area of JudgED.

CHAPTER 6

Discussion

This work presented a prototypical serious game to train intuitive decision-making processes of martial arts referees through immediate feedback. The prototype was designed, developed, and evaluated with the aim of answering the four defined research questions. The subsequent sections provide answers to these research questions by discussing them based on the results presented in this work.

6.1 Identified Requirements

Which requirements can be identified for a serious game to train decision-making skills of martial arts referees in terms of decision accuracy and reaction time?

Based on the core vision to build a video-based serious game to train the decision-making skills of martial arts referees, the requirements were collected through domain analysis and semi-structured interviews with six subject matter experts for kickboxing and karate Kumite. The initially gathered requirements were gradually refined in several iterations of exploratory prototyping and evolutionary system development.

The requirements engineering process identified the need to develop a serious game consisting of two modules: (a) a video-based training interface that allows referees to practice their decision-making skills by interactively judging a series of fight scenes and receiving immediate feedback on their correctness and (b) a content and administration interface that allows authorized referees to define, organize and evaluate the training sessions for the players of the serious game. A total of 25 requirements were identified, of which 13 relate to the content and administration module and nine relate to the serious game. The other three requirements relate to cross-functional concerns that can be mapped to both modules. The complete catalogue of requirements can be found in section 4.2.2.

Since the interviewees had backgrounds as referees, athletes, coaches, course instructors, and scientists, the collected requirements describe a prototype that can be used for scientific and educational purposes. This is particularly reflected in the requirements for the content and administration module. In addition to the basic requirements of defining video scenes, organizing them in playlists, and making them accessible to referees, this module also covers features such as managing users, organizing courses, and querying expressive statistics to track the performance of course participants. Considering ideas from these different perspectives made it possible to define comprehensive requirements for a serious game to train the decision-making skills of martial arts referees, which can be used both in scientific studies and in educational settings.

6.2 Performance Measurement

How can the serious game accurately determine the correctness and reaction time of the judgments entered by the martial arts referee based on the events appearing in the fight scene presented in the form of a streaming video?

Several considerations and techniques in both design and development contribute to accurately determining the correctness and reaction time of decisions made by martial arts referees in the serious game. This includes (i) precisely defining decisions on video scenes, (ii) pinpointing the time at which the user interacts with the video scene, and (iii) correlating the decisions entered by the user with the decisions defined on the video scene.

The video scenes presented to the referees in the serious game must be precisely annotated with decisions based on the events appearing in the fight scene. Each defined decision includes (i) the exact point in time to the nearest millisecond, (ii) the athlete (red or blue) to whom the decision is attributed, (iii) the technique that leads to the particular score or penalty, and (iv) the decision value. In addition to precisely defining video scenes, content creators were encouraged to only define decisions for events that were clearly visible in the footage. Each video scene was created by experienced kickboxing referees and athletes. Each defined decision was reviewed by an expert referee with more than 10 years of refereeing experience. The definition of high-quality content forms the basis to determine the correctness of the referees' decisions by comparing them with the expert-defined decisions in the video scene.

Utility functions of the client-side video player library were used to determine the point in time of the user's interaction in accordance with the progress of the streaming video. Since two different video player libraries were used for the web and mobile versions of the serious game, the quality of the time determined differs between these variants. The web version of the serious game actively queries the progress of the video with every decision made by the player, which allows for determining the exact time of the user's input. In contrast, the mobile version relies on asynchronous processing *timeupdate* events, which

are triggered at an interval of 250 ms. Regardless of the library used, this allows the conclusion that the serious game is able to record the player's decisions according to the progress of the streaming video, tolerating a maximum inaccuracy of 250 ms.

To compute the metrics of decision accuracy and reaction time, the decisions entered by the player are compared to the decisions defined in the respective video scene. Since it is not always unambiguous which player inputs were intended for which defined decision, an algorithm was developed that correlates player decisions with defined decisions using a defined set of rules. The main result of this algorithm is a list of tuples, each tuple containing the player's decision and the associated defined decision. Each decision's correctness and reaction time are determined by comparing the player's decision with the associated defined decision in the tuple. A comprehensive description of the data recording and measurement approach can be found in section 4.6.

6.3 Performance of Martial Arts Referees

How do professional martial arts referees perform in the serious game in terms of decision accuracy and reaction time, and how does this relate to their refereeing experience and general reaction time?

The decision accuracy and reaction time of licensed kickboxing referees in the serious game was evaluated in the form of a field experiment during the Austrian WAKO Kickboxing Championship 2022 in Graz. A total of 16 licensed kickboxing referees participated in the field experiment, which consisted of two tests performed in the serious game. In both tests, participants had to judge a series of video scenes as accurately and quickly as possible. During these tests, performance data was recorded in terms of decision accuracy and reaction time. In addition, a standardized test to measure the participants' general reaction times was conducted.

Consistent with hypothesis $H_{3.1}$, no significant differences were found between both tests for the metrics decision accuracy and reaction time. The overall results showed a mean decision accuracy of 43.011 % and a mean reaction time of 1.022 s. Thus, the results do not support the hypothesis $H_{3.2}$ stating that the average decision accuracy of referees in the serious game is greater than 50 %. Apart from the fact that the referees used the serious game for the first time, the low decision accuracy was partially caused by the conspicuously high error rate when judging events defined as penalties (i.e. warnings or exits). Only 4.475 % of all decisions defined as penalties were judged correctly. Excluding these decisions from the data set would increase the overall decision accuracy from 43.011 % to 49.626 %. Further examination of the decision accuracy reveals a large discrepancy between the performance of referees in the Tatami disciplines and Ring disciplines. While Tatami referees achieved a decision accuracy of 50.460 %, Ring referees only achieved a decision accuracy of 30.596 %. The difference could be explained by the fact that the average difficulty of video scenes in the Ring disciplines was 3.1 compared to 2.6 for the

Tatami disciplines. Moreover, video scenes of Ring disciplines contain several closely spaced decisions that are believed to be more difficult to judge. While video scenes from the Ring disciplines contained an average of 12.375 decisions, video scenes from the Tatami disciplines contained an average of only 2.288 decisions. Consistent with this observation and the hypothesis $H_{3.4}$, the results showed a significant correlation of the video scenes' difficulty rank with decision accuracy and reaction time. This also confirms the quality of the video scenes' difficulty rating assigned by the expert referee.

Besides the low decision accuracy, the overall level of agreement among the referees in the serious game on the judgments of expert-defined decisions turned out to be *poor*. Thus, the results do not support the hypothesis $H_{3.6}$ expecting a *moderate* level of agreement. The poor level of agreement is also reflected in the low probability of deriving a correct decision in a team of three independent referees, considering that the consent of at least two judges is required. While the probability of making a correct decision in the discipline *Point fighting* increases when judged within a team (from 52.963 % to 54.439 %), this is not the case for the other examined disciplines.

In addition to calculating the statistics described above, the performance data was analyzed for possible relationships with the referees' general reaction time and refereeing experience. While we hypothesized a correlation between these metrics, the results do not support the statement. Contrary to hypothesis $H_{3.5}$, no significant linear relationship was found between reaction time in JudgeED and reaction time measured by the four-choice reaction time tests. Moreover, the results show no significant correlation between refereeing experience and the performance data in JudgED for both decision accuracy and reaction time, which rejects the hypothesis $H_{3.3}$. This suggests that the results of the field experiment do not allow the conclusion that there is a correlation between in-game results and the performance of referees in real-life competitions, assuming more experienced referees have higher refereeing skills. To examine the existence of a potential relationship between in-game and on-field performance, the conduction of further studies is required. Table 6.1 lists the hypotheses for the research question RQ_3 and the information on whether they are supported or rejected.

Because no studies have examined martial arts referees' decision accuracy and reaction time, the results reported in this study provide the first performance data in this area. Therefore, the data can only be compared with studies of other sports. For example, Larkin et al. [LBDL11] examined the decision accuracy in the judgment of tackles in Australian football by using a video-based testing protocol. Depending on the expert level, referees reached a decision accuracy between 53.7 and 61.1 %. In another study, Mascarenhas et al. [MCMM05] examined a video-based training program to develop shared mental models of rugby referees. On average, the referees achieved a decision accuracy of 44.7 % for judging the tackles presented in the pre-test. However, compared to the study conducted in this work, a decision was only considered correct if both decision and justification were correct. Although these studies examine the decision accuracy of referees by utilizing video-based approaches, the different characteristics of the sports, programs and test setups do not allow a meaningful comparison.

Table 6.1: Supported and rejected hypotheses stated for research question RQ_3.

ID	Hypothesis	Support?
$H_{3.1}$	The decision accuracy of referees in the serious game does not improve between two consecutive tests performed in the serious game without feedback.	✓
$H_{3.2}$	The average decision accuracy of referees in the serious game is greater than 50 %.	✗
$H_{3.3}$	The level of refereeing experience has a positive impact on the decision accuracy in the serious game.	✗
$H_{3.4}$	Video scenes with increased difficulty rank show a reduced decision accuracy in the serious game.	✓
$H_{3.5}$	The general reaction time of referees has a positive impact on the reaction time in the serious game.	✗
$H_{3.6}$	The level of agreement among the referees in the serious game on the judgments of expert-defined decisions is *moderate* [KL16].	✗

6.4 Potential of JudgED for Referee Training

How do professional martial arts referees assess the potential of the serious game to enhance referee training?

The potential of JudgED to enhance referee training was examined by a survey conducted after the field experiment. Among other topics, the participating referees were asked how they assess the potential, target groups, and possible application areas of JudgED. In general, respondents believe that JudgED has the potential to improve referee education classes. They believe it is a useful tool to train decision-making skills for referees of all experience levels. In addition to using JudgeED in courses and seminars, it was also seen as a complementary training method that can be used at home. Most respondents would use JudgeED on a tablet device.

Besides the current purpose of JudgED to serve as a prototype for scientific purposes, it was also designed to be used in non-scientific settings such as referee courses and seminars. The features of the content and administration module already cover use cases to manage users, organize courses, and query comprehensive statistics to track the performance of course participants. In addition, the system already has a specific user role *course organizer*, whose permissions are tailored to the administration of courses. This allows course instructors to prepare selective training videos and playlists focused on specific didactic aspects and make them available to course participants. Course participants can use the video scenes included in these playlists to train their decision-making skills and compare their performance with other referees in the course. The serious game has thus already reached a level of maturity to be used as a practical intervention in educational settings.

6.5 Limitations

The study's main limitation concerns the experimental design for evaluating referees' performance in the serious game. Besides the relatively small number of subjects who participated in the field experiment, the short period to become familiar with the mechanics of the serious game might have contributed to the low performance in decision accuracy. Moreover, the study is limited by the number of video scenes used to assess the referees' performance during the field experiment. Although the video scenes used for the field experiment are of high quality, as they were carefully selected and annotated by an expert referee, the limited resources for conducting the study did not allow a larger volume to be produced. In addition, the viewing perspective may not represent the natural environment of competitive decision-making. On the one hand, the video scenes are not recorded from the first-person perspective, as they mainly consist of footage from real-life competitions. On the other hand, the field experiment was conducted on 10-inch tablets, which may have made it difficult to spot certain decisions. Because the field experiment and survey exclusively focused on kickboxing referees, the results obtained are not transferrable to martial arts in general. Therefore, obtained results regarding referee performance in the serious game and the potential of JudgeED to improve referee education cannot be generalized to all martial arts.

In addition to the limitations regarding the experimental design, the expressiveness of the study is weakened by the small number of martial arts representatives in the analysis and design of the serious game, which does not allow a generalisation of the collected requirements and the designed serious game for sports other than kickboxing and karate Kumite. While the serious game was designed to flexibly add new martial arts disciplines to JudgeED through customisation of configuration files and minimal code changes, essential characteristics of sports other than kickboxing and karate Kumite might not have been anticipated. Therefore, the generalisation of JudgED as a tool for training martial arts referee decision-making processes may not be valid.

CHAPTER 7

Conclusion and Future Work

While participation in sports competitions is recognised as the ideal method for developing decision-making skills, the limited number of competitive events throughout the year does not allow sufficient training hours to reach expert level. Using video-based training programs is a possible solution to compensate for the lack of training intensity. This work presents a video-based serious game to train decision-making processes of martial arts referees through immediate feedback. While few studies have examined systems to train the decision-making skills of referees, the serious game developed in this work is the first program specifically focused on martial arts referees.

The prototype was designed based on principles of social judgment theory, which states that decisions are derived by traversing the steps of perception, categorisation, memory processing, and information integration. While all steps are essential to derive a decision, the serious game in this work focuses on the step of categorisation, where perceived information is compared to knowledge stored in memory. The concept of multiple-cue probability learning is applied to train the categorisation of perceived information, which facilitates skills based on repeated exposure to probabilistic information. The prototype embodies this concept by asking referees to judge a series of video scenes depicting fight situations. The feedback on the correctness of the entered decisions aims to establish cue-outcome relations that train the decision accuracy of referees.

With these theoretical considerations in mind, the system was developed in ten iterations comprising analysis, design, and development activities. In each iteration, the feedback from subject matter experts in kickboxing and karate Kumite was incorporated accordingly. This allowed to gradually improve the system until all requirements were fulfilled and the prototype reached maturity to be evaluated in a field experiment. The final prototype consisted of two modules: (a) a training module to train decision-making skills of martial arts referees through a video-based approach utilising immediate feedback and measuring the performance of referees in terms of decision accuracy and reaction

time and (b) a content and administration module allowing authorised referees to define video scenes used in the serious game, organise training sessions, and analyse the players' performance. The conducted field experiment showed that the overall decision accuracy of licensed kickboxing referees in the serious game was below 50 %. Remarkably, misjudgments when assessing penalties contributed significantly to the low decision accuracy. Contrary to other studies, experienced referees did not outperform less experienced ones in the serious game. In addition, a low general reaction time of the referees did not lead to better results in the serious game. To confirm these results, further studies with more participants are needed to demonstrate external validity.

The present study did not evaluate the prototype in terms of effectiveness and motivation, which are essential characteristics of a serious game. To test these properties, the conduction of a comprehensive study is suggested. The effectiveness of the serious game in improving decision-making skills can be tested by conducting a field experiment in the form of a *pretest-posttest control group design* [CBL14]. To test the ability of the serious game to intrinsically motivate referees, the implementation of a questionnaire comprising standardised questions from the *Post-Experimental Intrinsic Motivation Inventory scale* (IMI) [inta] is suggested as presented in Appendix 7. Salmhofer et al. [SGFH+22] proposes a comprehensive setup for evaluating the serious game in terms of effectiveness and motivation, which can serve as a guideline for future studies. Although such a study does not allow a conclusion as to whether improvements in the serious game can be transferred to real competitions, it does examine whether the promising results from related studies on video-based decision-making training programs can be transferred to the sport of martial arts.

During interviews and informal feedback from referees, it turned out that the developed prototype was not perceived as a serious game but rather as training software. This impression is underpinned by the results of the survey, the answers to which indicate a low willingness to train with the serious game voluntarily. This lack of intrinsic motivation can be caused by the small number of game elements incorporated in the serious game. Although the prototype contains rudimentary game elements, such as a leaderboard and the current rank, there is still room for improvement in this area. In order to increase the intrinsic motivation of referees to train in the serious game, performance-related badges and levels could be introduced. In addition, automatic difficulty adjustment based on the referee's performance could be introduced to continuously challenge them and maintain a flow state where the game fully engages them.

The serious game thrives on the definition of high-quality video scenes appropriately annotated with decisions. Therefore, the developed prototype offers not only a serious game for training decision-making skills but also a content and administration module that provides expert referees with an easy way to create video scenes that can be used in the serious game without depending on third-party software. Feedback from referees indicates the potential of JudgED as a useful tool to train the decision-making skills of martial arts referees. The complementary use of JudgED allows for accumulating

practical training intensity, which would hardly be possible by solely judging real-world competitions.

List of Figures

List of Tables

Bibliography

[Abt70] Clark C. Abt. *Serious Games*. NY: Viking Press, 1970.

[act] Active record basics - ruby on rails guides. `https://guides.rubyonrails.org/active_record_basics.html`. [Online; last accessed 2022-09-29].

[Ada15] William C. Adams. Conducting Semi-Structured Interviews. In *Handbook of Practical Program Evaluation*, pages 492–505. John Wiley & Sons, Ltd, 2015.

[ALS18] Theofylaktos Anastasiadis, Georgios Lampropoulos, and Kerstin Siakas. Digital Game-based Learning and Serious Games in Education. *International Journal of Advances in Scientific Research and Engineering*, 4(12):139–144, 2018.

[anga] Angular - components overview. `https://angular.io/guide/component-overview`. [Online; last accessed 2022-10-06].

[angb] Angular - template syntax. `https://angular.io/guide/template-syntax`. [Online; last accessed 2022-10-06].

[angc] Angular - the modern web developer's platform. `https://angular.io/`. [Online; last accessed 2022-09-24].

[angd] Angular - what is angular? `https://angular.io/guide/what-is-angular`. [Online; last accessed 2022-10-06].

[AT09] Shane Allua and Cheryl Bagley Thompson. Inferential statistics. *Air medical journal*, 28(4):168–171, 2009.

[BB10] Johannes Breuer and Gary Bente. Why so serious? On the relation of serious games and learning. *Journal for Computer Game Culture*, 4 (1):7–24, 2010.

[BCC+17] Ioana Baldini, Paul Castro, Kerry Chang, Perry Cheng, Stephen Fink, Vatche Ishakian, Nick Mitchell, Vinod Muthusamy, Rodric Rabbah, Aleksander Slominski, and Philippe Suter. Serverless Computing: Current

Trends and Open Problems. In *Research Advances in Cloud Computing*, pages 1–20. Springer Singapore, 2017.

[BFS04] Herbert Bless, Klaus Fiedler, and Fritz Strack. *Social cognition: How individuals construct social reality.* Psychology Press, 2004.

[BKKZ90] Reinhard Budde, Karlheinz Kautz, Karin Kuhlenkamp, and Heinz Züllighoven. What is prototyping? *Information Technology & People*, 6(2):89–95, 1990.

[BL07] R Bender and S Lange. Was ist der p -Wert? *DMW - Deutsche Medizinische Wochenschrift*, 132(1):15–16, 2007.

[BLM09] Michel Beaudouin-Lafon and Wendy E Mackay. Prototyping tools and techniques. In *Human-Computer Interaction*, pages 137–160. CRC Press, 2009.

[BPS09] Ralf Brand, Henning Plessner, and Geoffrey Schweizer. Conceptual considerations about the development of a decision-making training method for expert soccer referees. In *Perspectives on cognition and action in sport*, pages 181–190. Hauppauge, NY: Nova Science, 2009.

[Bro96] John Brooke. Sus - a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[Bru52] Egon Brunswik. *The conceptual framework of psychology. (Int. Encycl. unified Sci., v. 1, no. 10.).* Univ. Chicago Press, 1952.

[BSW14] Josette Bettany-Saltikov and Victoria Jane Whittaker. Selecting the most appropriate inferential statistical test for your quantitative research study. *Journal of Clinical Nursing*, 23(11-12):1520–1531, 2014.

[CBKPC20] Tomas Carlsson, Jozef Berglez, Sebastian Koivisto Persson, and Magnus Carlsson. The impact of video review in karate kumite during a Premier League competition. *International Journal of Performance Analysis in Sport*, 20(5):846–856, 2020.

[CBL14] William D Crano, Marilynn B Brewer, and Andrew Lac. Designing Experiments - Variations on Basics. In *Principles and methods of social research*, pages 83–100. Routledge, 2014.

[CGGA+10] Peter Catteeuw, Bart Gilis, José-María García-Aranda, Fernando Tresaco, Johan Wagemans, and Werner Helsen. Offside decision making in the 2002 and 2006 FIFA World Cups. *Journal of sports sciences*, 28(10):1027–1032, 2010.

[CRSH17] Seung Youn Chyung, Katherine Roberts, Ieva Swanson, and Andrea Hankinson. Evidence-based survey design: The use of a midpoint on the likert scale. *Performance Improvement*, 56(10):15–23, 2017.

[Csi90] Mihaly Csikszentmihalyi. *Flow: The psychology of optimal experience*. Harper & Row New York, 1990.

[CV97] Mahil Carr and June Verner. Prototyping and software development approaches. *Department of Information Systems, City University of Hong Kong*, pages 319–338, 1997.

[CVS16] Jan Carboch, Katerina Vejvodova, and Vladimir Suss. Analysis of errors made by line umpires on ATP tournaments. *International Journal of Performance Analysis in Sport*, 16(1):264–275, 2016.

[DCM+07] Elizabeth Dean, Rachel Caspar, Georgina McAvinchey, Leticia Reed, and Rosanna Quiroz. Developing a low-cost technique for parallel cross-cultural instrument development: The question appraisal system (qas-04). *International Journal of Social Research Methodology*, 10(3):227–241, 2007.

[DELG85] ED Diener, Robert A Emmons, Randy J Larsen, and Sharon Griffin. The satisfaction with life scale. *Journal of personality assessment*, 49(1):71–75, 1985.

[DGEW16] Ralf Dörner, Stefan Göbel, Wolfgang Effelsberg, and Josef Wiemeyer. *Serious Games: Foundations, Concepts and Practice*, chapter Introduction, pages 1–34. Springer International Publishing, 2016.

[djaa] Django - django at a glance. `https://docs.djangoproject.com/en/4.1/intro/overview/`. [Online; last accessed 2022-09-28].

[djab] Django - the web framework for perfectionists with deadlines. `https://www.djangoproject.com/`. [Online; last accessed 2022-09-24].

[djac] Django faqs: Mtv (model - template - view). `https://docs.djangoproject.com/en/4.1/faq/general/`. [Online; last accessed 2022-09-28].

[djad] The django template language. `https://docs.djangoproject.com/en/4.1/ref/templates/language/`. [Online; last accessed 2022-09-28].

[DoLGSR] Sport Department of Local Government, Cultural Industries – Sport, and Recreation. Officiating in sport and recreation. `https://www.dlgsc.wa.gov.au/sport-and-recreation/People-and-industry/officiating`. [Online; last accessed 2022-07-07].

[Eis15] Bonnie Eisenman. *Learning react native: Building native mobile apps with JavaScript*. O'Reilly Media, Inc., 2015.

[EKTR93] K. Anders Ericsson, Ralf T. Krampe, and Clemens Tesch-Römer. The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3):363–406, 1993.

[emb] Vimeo content delivery network - embed presets. `https://developer.vimeo.com/api/reference/embed-presets`. [Online; last accessed 2022-09-16].

[EN11] Simon Egenfeldt-Nielsen. *Beyond edutainment: Exploring the educational potential of computer games*. Lulu.com, 2011.

[expa] Express - fast, unopinionated, minimalist web framework for node.js. `https://expressjs.com/`. [Online; last accessed 2022-09-16].

[expb] Express middleware. `https://expressjs.com/en/guide/using-middleware.html`. [Online; last accessed 2022-09-16].

[expc] Express routing. `https://expressjs.com/en/guide/routing.html`. [Online; last accessed 2022-09-16].

[expd] Template engines compatible with express js. `https://expressjs.com/en/resources/template-engines.html`. [Online; last accessed 2022-09-24].

[FG88] Mark G. Frank and Thomas Gilovich. The dark side of self- and social perception: Black uniforms and aggression in professional sports. *Journal of Personality and Social Psychology*, 54(1):74–85, 1988.

[fig] Fight trainer - mixed martial arts. `https://play.google.com/store/apps/details?id=com.criticalbit.fightingtrainer`. [Online; last accessed 2021-06-08].

[fira] Google firebase admin node.js sdk - source code repository. `https://github.com/firebase/firebase-admin-node`. [Online; last accessed 2022-09-16].

[firb] Google firebase authentication. `https://firebase.google.com/docs/auth`. [Online; last accessed 2022-09-16].

[firc] Google firebase client web sdk - source code repository. `https://github.com/firebase/firebase-js-sdk`. [Online; last accessed 2022-09-16].

[ird] Google firebase platform. `https://firebase.google.com/`. [Online; last accessed 2022-09-16].

[fire] Google firebase pricing. https://firebase.google.com/pricing. [Online; last accessed 2022-09-16].

[Flo84] Christiane Floyd. A Systematic Look at Prototyping. In *Approaches to Prototyping*, pages 1–18. Springer Berlin Heidelberg, 1984.

[FM09] Murray J. Fisher and Andrea P. Marshall. Understanding descriptive statistics. *Australian Critical Care*, 22(2):93–97, 2009.

[fou] Inquisit - four choice reaction time task. https://www.millisecond.com/download/library/fourchoicereactiontimetask. [Online; last accessed 2022-11-11].

[Fow] Martin Fowler. Active record pattern. https://www.martinfowler.com/eaaCatalog/activeRecord.html. [Online; last accessed 2022-09-29].

[FT91] Susan T. Fiske and Shelley E. Taylor. *Social cognition, 2nd ed.* Mcgraw-Hill Book Company, 1991.

[GB08] Andreas Glöckner and Tilmann Betsch. Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making*, 3:215–228, 2008.

[get] Get it right - platform for referees. https://www.getitrighttraining.com/. [Online; last accessed 2021-05-19].

[git] Gitlab - the devsecops platform. https://about.gitlab.com/. [Online; last accessed 2022-09-16].

[GM20] Emily Geisen and Joe Murphy. *A Compendium of Web and Mobile Survey Pretesting Methods*, chapter 12, pages 287–314. John Wiley & Sons, 2020.

[GMR07] Rainer Göb, Christopher McCollin, and Maria Fernanda Ramalhoto. Ordinal Methodology in the Analysis of Likert Scales. *Quality & Quantity*, 41(5):601–626, 2007.

[GSM02] Mark A. Gluck, Daphna Shohamy, and Catherine Myers. How do People Solve the "Weather Prediction" Task?: Individual Variability in Strategies for Probabilistic Category Learning. *Learning & Memory*, 9(6):408–418, 2002.

[GW16] Stefan Göbel and Viktor Wendel. Personalization and Adaptation. In Ralf Dörner, Stefan Göbel, Wolfgang Effelsberg, and Josef Wiemeyer, editors, *Serious Games*, pages 161–210. Springer International Publishing, Cham, 2016.

[GY16] Ulas Gulec and Murat Yilmaz. A serious game for improving the decision making skills and knowledge levels of Turkish football referees according to the laws of the game. *SpringerPlus*, 5(1):1–10, 2016.

[hera] Heroku - cloud application platform. `https://www.heroku.com/`. [Online; last accessed 2022-09-16].

[herb] Heroku dynos - lightweight containers for running apps. `https://www.heroku.com/dynos`. [Online; last accessed 2022-09-16].

[Hig96] Edward Tory Higgins. Knowledge activation: Accessibility, applicability, and salience. In *Social psychology: Handbook of basic principles.*, pages 133–168. The Guilford Press, 1996.

[HMPR04] Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. Design Science in Information Systems Research. *MIS Quarterly*, 28(1):75–105, 2004.

[Hog11] Robin M Hogarth. On the learning of intuition. In *Intuition in judgment and decision making*, pages 111–126. Psychology Press, 2011.

[inq] Inquisit player app. `https://www.millisecond.com/download/inquisitweb6`. [Online; last accessed 2022-11-11].

[inta] Intrinsic motivation inventory (imi). `https://selfdeterminationtheory.org/intrinsic-motivation-inventory/`. [Online; last accessed 2022-04-01].

[intb] React - internationalization library. `https://formatjs.io/docs/react-intl/`. [Online; last accessed 2022-09-16].

[jit] Jitsi meet. `https://meet.jit.si/`. [Online; last accessed 2022-11-10].

[JSX] React jsx - javascript xml. `https://www.w3schools.com/react/react_jsx.asp`. [Online; last accessed 2022-09-16].

[KB13] Philip T. Kortum and Aaron Bangor. Usability Ratings for Everyday Products Measured With the System Usability Scale. *International Journal of Human–Computer Interaction*, 29(2):67–76, 2013.

[KCL+21] Aden Kittel, Ian Cunningham, Paul Larkin, Matthew Hawkey, and Geraldine Rix-Lièvre. Decision-making training in sporting officials: Past, present and future. *Psychology of Sport and Exercise*, 56:102003, 2021.

[KL16] Terry K. Koo and Mae Y. Li. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016.

[KP09] Jon Krosnick and Stanley Presser. Question and Questionnaire Design. In *Handbook of Survey Research*, pages 263—-313. 2009.

[KS10] Henrik Kniberg and Mattias Skarin. *Kanban and Scrum-making the most of both*. Lulu. com, 2010.

[KSG94] B J Knowlton, L R Squire, and M A Gluck. Probabilistic classification learning in amnesia. *Learning & Memory*, 1(2):106–120, 1994.

[KSR17] Stephanie A Kovalchik, Jeff Sackmann, and Machar Reid. Player, official or machine?: uses of the challenge system in professional tennis. *International Journal of Performance Analysis in Sport*, 17(6):961–969, 2017.

[LBDL11] Paul Larkin, Jason Berry, Brian Dawson, and Brendan Lay. Perceptual and decision-making skills of Australian football umpires. *International Journal of Performance Analysis in Sport*, 11(3):427–437, 2011.

[LEES14] Fedwa Laamarti, Mohamad Eid, and Abdulmotaleb El Saddik. An Overview of Serious Games. *International Journal of Computer Games Technology*, 2014:1–15, 2014.

[Lik32] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[LMB+18] Paul Larkin, Christopher Mesagno, Jason Berry, Michael Spittle, and Jack Harvey. Video-based training to improve perceptual-cognitive decision-making performance of Australian football umpires. *Journal of sports sciences*, 36(3):239–246, 2018.

[LNKS06] David A. Lagnado, Ben R. Newell, Steven Kahan, and David R. Shanks. Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, 135(2):162–183, 2006.

[ltia] Learning tools interoperability. `https://www.imsglobal.org/activity/learning-tools-interoperability`. [Online; last accessed 2022-09-16].

[ltib] Ltijs - easily turn your web application into a lti 1.3 learning tool. `https://cvmcosta.github.io/ltijs`. [Online; last accessed 2022-09-16].

[Mac15] Clare MacMahon. *Sports officials and officiating: science and practice*. Routledge, 2015.

[Man11a] S Manikandan. Measures of central tendency: Median and mode. *Journal of Pharmacology and Pharmacotherapeutics*, 2(3):214–215, 2011.

[Man11b] S Manikandan. Measures of central tendency: The mean. *Journal of Pharmacology and Pharmacotherapeutics*, 2(2):140–142, 2011.

[mar] Martial codex - technique learning tool for athletes. https://www.martialcodex.com/. [Online; last accessed 2021-06-08].

[Mat08] George Mather. Perceptual uncertainty and line-call challenges in professional tennis. *Proceedings of the Royal Society B: Biological Sciences*, 275(1643):1645–1651, 2008.

[May15] Philipp Mayring. *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. Beltz, 2015.

[MC05] David R Michael and Sandra L Chen. *Serious games: Games that educate, train, and inform*. Muska & Lipman/Premier-Trade, 2005.

[MCMM05] Duncan RD Mascarenhas, Dave Collins, Patrick W Mortimer, and Bob Morris. Training accurate and coherent decision making in rugby union referees. *The Sport Psychologist*, 19(2):131–147, 2005.

[MDKR12] Dietrich Albert Michael D. Kickmeier-Rust. *An alien's guide to multi-adaptive educational computer games*. Informing Science, 2012.

[MHSW07] Clare MacMahon, Werner F. Helsen, Janet L. Starkes, and Matthew Weston. Decision-making skills and deliberate practice in elite association football referees. *Journal of Sports Sciences*, 25(1):65–78, 2007.

[mona] Getting started with atlas and the mongodb query language (mql). https://www.mongodb.com/developer/products/atlas/getting-started-atlas-mongodb-query-language-mql/. [Online; last accessed 2022-09-16].

[monb] Mongodb - the developer data platform. https://www.mongodb.com. [Online; last accessed 2022-09-16].

[monc] Mongodb - what is mongodb? https://www.mongodb.com/docs/manual/core/databases-and-collections/. [Online; last accessed 2022-09-16].

[mond] Mongodb pricing. https://www.mongodb.com/pricing. [Online; last accessed 2022-09-16].

[mone] Mongodb wire protocol. https://www.mongodb.com/docs/manual/reference/mongodb-wire-protocol/. [Online; last accessed 2022-09-16].

[ms] Microsoft teams - video conferencing, meetings, calling. https://www.microsoft.com/de-at/microsoft-teams/group-chat-software/. [Online; last accessed 2022-11-10].

[MS14] Clare MacMahon and Bernd Strauß. The psychology of decision making in sport officals. In *An introduction to sport and exercise psychology*, pages 223–235. London: Routledge, 2014.

[mui] React - material ui library. `https://mui.com/`. [Online; last accessed 2022-09-16].

[MvD07] Ali Mesbah and Arie van Deursen. Migrating Multi-page Web Applications to Single-page AJAX Interfaces. In *11th European Conference on Software Maintenance and Reengineering (CSMR'07)*, pages 181–190, 2007.

[mys] Mysql database. `https://www.mysql.com/de/`. [Online; last accessed 2022-10-09].

[NBMW02] A.M Nevill, N.J Balmer, and A Mark Williams. The influence of crowd noise and experience upon refereeing decisions in football. *Psychology of Sport and Exercise*, 3(4):261–272, 2002.

[NHWA15] Kathryn E Newcomer, Harry P Hatry, Joseph S Wholey, and William C Adams. Conducting semi-structured interviews. In *Handbook of practical program evaluation*, volume 492, pages 492–505. 2015.

[Nie93] Jakob Nielsen. *Usability engineering*. Morgan Kaufmann, 1993.

[nod] Nodejs - run javascript everywhere. `https://nodejs.dev/`. [Online; last accessed 2022-09-16].

[npm] Node package manager (npm). `https://www.npmjs.com/`. [Online; last accessed 2022-09-16].

[NR81] Allen Newell and Paul S Rosenbloom. Mechanisms of skill acquisition and the law of practice. *Cognitive skills and their acquisition*, 1:1–55, 1981.

[oau] Open authorization - oauth 2.0. `https://oauth.net/2/`. [Online; last accessed 2022-09-16].

[OVB+00] Raôul RD Oudejans, Raymond Verheijen, Frank C Bakker, Jeroen C Gerrits, Marten Steinbrückner, and Peter J Beek. Errors in judging offside in football. *Nature*, 404(6773):33–33, 2000.

[Ove07] Hagen Overdick. The Resource-Oriented Architecture. In *2007 IEEE Congress on Services (Services 2007)*, pages 340–347. IEEE, 2007.

[PCY14] Debra Paul, James Cadle, and Donald Yeates. *Business analysis*. BCS, The Chartered Institute for IT, 2014.

[Per15] Cyril Pernet. Null hypothesis significance testing: a short tutorial. *F1000Research*, 4:621, 2015.

[PH06] Henning Plessner and Thomas Haar. Sports performance judgments from a social cognitive perspective. *Psychology of Sport and Exercise*, 7(6):555–575, 2006.

[PKL97] Roderick Perkins, Dan Smith Keller, and Frank Ludolph. Inventing the Lisa User Interface. *Interactions*, 4(1):40–53, 1997.

[pos] Postgresql - the world's most advanced open source relational database. `https://www.postgresql.org/`. [Online; last accessed 2022-10-09].

[PR15] Klaus Pohl and Chris Rupp. *Requirements engineering fundamentals: a study guide for the certified professional for requirements engineering exam, foundation level, IREB compliant*. Rocky Nook, 2015.

[pri] Prisma - next-generation node.js and typescript orm. `https://www.prisma.io/`. [Online; last accessed 2022-09-29].

[PWS+16] Koen Put, Johan Wagemans, Jochim Spitz, A Mark Williams, and Werner F Helsen. Using web-based training to enhance perceptual-cognitive skills in complex dynamic offside events. *Journal of Sports Sciences*, 34(2):181–189, 2016.

[PZB88] A Parsu Parasuraman, Valarie Zeithaml, and Leonard Berry. Servqual: A multiple-item scale for measuring consumer perceptions of service quality. *1988*, 64(1):12–40, 1988.

[QC16] Meihua Qian and Karen R. Clark. Game-based Learning and 21st century skills: A review of recent research. *Computers in Human Behavior*, 63:50–58, 2016.

[raia] Ruby on rails - a web-app framework that includes everything needed to create database-backed web applications according to the model-view-controller (mvc) pattern. `https://rubyonrails.org/`. [Online; last accessed 2022-09-29].

[raib] Ruby on rails - getting started with rails. `https://guides.rubyonrails.org/getting_started.html`. [Online; last accessed 2022-09-29].

[raic] Ruby on rails - the rails command line. `https://guides.rubyonrails.org/command_line.html`. [Online; last accessed 2022-09-29].

[reaa] React - a javascript library for building user interfaces. `https://reactjs.org/`. [Online; last accessed 2022-09-16].

[reab] React - building your own hooks. `https://reactjs.org/docs/hooks-custom.html`. [Online; last accessed 2022-09-16].

[reac] React - components and props. https://reactjs.org/docs/components-and-props.html. [Online; last accessed 2022-09-16].

[read] React native - javascript library for building mobile user interfaces. https://reactnative.dev/. [Online; last accessed 2022-09-16].

[rou] React - router library. https://github.com/remix-run/react-router. [Online; last accessed 2022-09-16].

[RR09] Rabindra Ratan and Ute Ritterfeld. *Serious games: Mechanisms and effects*, chapter 2, pages 10–24. Routledge, 2009.

[RSI96] Jim Rudd, Ken Stern, and Scott Isensee. Low vs. high-fidelity prototyping debate. *Interactions*, 3(1):76–85, 1996.

[SAZ+18] Kosovare Sahatqija, Jaumin Ajdari, Xhemal Zenuni, Bujar Raufi, and Florije Ismaili. Comparison between relational and NOSQL databases. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 216–221, 2018.

[SB98] Raymond Stefani and J. Bennett. Predicting outcomes. *Statistics in Sport*, pages 249–275, January 1998.

[SB16] Kyle Schutt and Osman Balci. Cloud software development platforms: A comparative overview. In *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 3–13, 2016.

[SCFY96] R.S. Sandhu, E.J. Coyne, H.L. Feinstein, and C.E. Youman. Role-based access control models. *Computer*, 29(2):38–47, 1996.

[Sco15] Emmit Scott. *SPA Design and Architecture: Understanding Single Page Web Applications*. Manning Publications Co., 1st edition, 2015.

[seq] Sequelize - feature-rich orm for modern typescript and javascript. https://sequelize.org/. [Online; last accessed 2022-09-29].

[SGFH+22] André Salmhofer, Lucian Gutica-Florescu, Dominik Hoelbling, Roland Breiteneder, Rene Baranyi, and Thomas Grechenig. 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 - icSPORTS*, pages 29–40. SciTePress, 2022.

[She83] Paul B. Sheatsley. Questionnaire Construction and Item Writing. In *Handbook of survey research*, Quantitative studies in social relations, pages 195–230. Acad. Press, 1983.

[SHM09] Jeff Sinclair, Philip Hingston, and Martin Masek. Exergame development using the dual flow model. In *Proceedings of the Sixth Australasian Conference on Interactive Entertainment - IE '09*, pages 1–7. ACM Press, 2009.

[SJB07] Tarja Susi, Mikael Johannesson, and Per Backlund. Serious games: An overview. pages 1–26, 2007.

[SPKB11] Geoffrey Schweizer, Henning Plessner, Daniela Kahlert, and Ralf Brand. A video-based training method for improving soccer referees' intuitive decision-making skills. *Journal of Applied Sport Psychology*, 23(4):429–442, 2011.

[spo] Sportdata event technology. `https://www.sportdata.org/`. [Online; last accessed 2022-09-16].

[sps] Ibm spss statistics. `https://www.ibm.com/at-de/products/spss-statistics`. [Online; last accessed 2022-11-11].

[sql] Sqlite database. `https://www.sqlite.org/`. [Online; last accessed 2022-10-09].

[SS17] Kalpna Sagar and Anju Saha. A systematic review of software usability studies. *International Journal of Information Technology*, 2017.

[staa] Stack overflow developer survey 2022. `https://survey.stackoverflow.co/2022/`. [Online; last accessed 2022-10-07].

[stab] Stack overflow trends. `https://insights.stackoverflow.com/trends`. [Online; last accessed 2022-10-07].

[Sta20] Katarzyna Stapor. Descriptive and Inferential Statistics. In *Introduction to Probabilistic and Statistical Methods with Examples in R*, pages 63–131. Springer International Publishing, 2020.

[Sub19] Vasan Subramanian. *Pro MERN Stack: Full Stack Web App Development with Mongo, Express, React, and Node*. Apress, 2019.

[Sve01] Elisabeth Svensson. Guidelines to statistical evaluation of data from rating scales and questionnaires. *Journal of rehabilitation medicine: official journal of the UEMS European Board of Physical and Rehabilitation Medicine*, 33:47–8, 2001.

[SW05] Penelope Sweetser and Peta Wyeth. GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment*, 3(3):1–24, 2005.

[SWM+21] Jochim Spitz, Johan Wagemans, Daniel Memmert, A Mark Williams, and Werner F Helsen. Video assistant referees (VAR): The impact of technology on decision making in association football referees. *Journal of Sports Sciences*, 39(2):147–153, 2021.

[tus] tus - open protocol for resumable file uploads. https://tus.io/. [Online; last accessed 2022-09-16].

[vC17] Wilfred van Casteren. The Waterfall Model and the Agile Methodologies: A comparison by project characteristics. 2017.

[vima] Vimeo content delivery network. https://vimeo.com/. [Online; last accessed 2022-09-16].

[vimb] Vimeo content delivery network - privacy settings. https://vimeo.com/blog/post/vimeo-video-settings-upload/. [Online; last accessed 2022-09-16].

[vimc] Vimeo content delivery network - scopes. https://developer.vimeo.com/api/authentication. [Online; last accessed 2022-09-16].

[vimd] Vimeo developer api. https://developer.vimeo.com/api/. [Online; last accessed 2022-09-16].

[Vog21] Patric U. B Vogel. *Laborstatistik für technische Assistenten und Studierende*. essentials. Springer Fachmedien Wiesbaden Imprint: Springer Spektrum, 1st edition, 2021.

[VS11] Irena Petrijevcanin Vuksanovic and Bojan Sudarevic. Use of web application frameworks in the development of small applications. In *2011 Proceedings of the 34th International Convention MIPRO*, pages 458–462, 2011.

[vuea] Vue - creating a vue application. https://vuejs.org/guide/essentials/application.html. [Online; last accessed 2022-10-06].

[vueb] Vue - introduction. https://vuejs.org/guide/introduction.html. [Online; last accessed 2022-10-06].

[vuec] Vue - the progressive javascript framework. https://vuejs.org/. [Online; last accessed 2022-10-06].

[Vya22] Rishi Vyas. Comparative Analysis on Front-End Frameworks for Web Applications. *International Journal for Research in Applied Science and Engineering Technology*, 10(7):298–307, 2022.

[waka] Wako referee rules. https://wako.sport/rules/. [Online; last accessed 2023-05-13].

[wakb] Wako rules live - rule learning tool for kickboxing referees. https://www.facebook.com/pulsebysport2016/. [Online; last accessed 2021-06-08].

[WCS10] Bert Weijters, Elke Cabooter, and Niels Schillewaert. The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27(3):236–247, 2010.

[WCT88] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[WL99] Gordon Willis and Judith Lessler. Question appraisal system qas-99. 1999.

[WNM6] Josef Wiemeyer, Lennart Nacke, Christiane Moser, and Florian 'Floyd' Mueller. Player Experience. In *Serious Games*, pages 243–271. Springer International Publishing, 2016.

[XLKR00] Xia Cai, M.R. Lyu, Kam-Fai Wong, and Roy Ko. Component-based software engineering: technologies, development frameworks, and quality assurance schemes. In *Proceedings Seventh Asia-Pacific Software Engeering Conference. APSEC 2000*, pages 372–379. IEEE Comput. Soc, 2000.

[ZC05] Didar Zowghi and Chad Coulin. Requirements Elicitation: A Survey of Techniques, Approaches, and Tools. In *Engineering and Managing Software Requirements*, pages 19–46. Springer Berlin Heidelberg, 2005.

[zoo] Zooom app - platform for referees. `https://zooomapp.com/`. [Online; last accessed 2021-05-19].

Questionnaire 1

This questionnaire was designed to gather feedback from participants after the field experiment. It covers questions to gather (i) basic demographic data, experience, and opinions about computer-supported referee education, (ii) feedback on certain topics relevant for future feature development, (iii) the self-assessment of decision-making abilities, and (iv) opinions about potential application areas and target groups of the serious game. Furthermore, it enables further insights through the targeted combination of questionnaire results with performance data recorded in the field experiment.

The questionnaire starts with an introduction, informing the participants about the general purpose of the questionnaire. Although the questionnaire requires the respondents to be already informed about the serious game examined in the study, the introduction section includes a short video summarizing the purpose and mechanics of *JudgED*. The participants are prompted to watch the video before starting with the questionnaire. In addition, for privacy reasons, the introduction includes information about how the participants' data will be processed.

The actual questionnaire comprises 58 questions structured in 11 sections. Chapters are compiled in a logical order and contain coherent questions. Questions within the chapters are arranged to query basic information before going into more detail. As the initial target audience included Austrian kickboxing referees, the questionnaire was written in German. The following paragraph enumerates the sections and compactly describes their content:

- **Basics:** General information of the respondents concerning age and gender, as well as refereeing background, including formal education and hands-on experience.
- **Computer skills and usage habits:** Basic computer skills, usage habits, computer access, and experience with computer-based referee training.
- **Computer-supported referee education:** The potential of using digital media and computer-based systems in referee education to enhance rule knowledge and decision-making skills.
- **Competition in education:** The contribution of competitive elements such as comparisons and rankings in referee education and its effect on fun and motivation.

- **Importance of metrics:** Expressiveness of performance-related metrics, their relevance to be displayed in JudgED, and their adequacy to compare referees in JudgED.
- **Rating of feature variants:** Feedback about the implementation of certain features with respect to the scope of rankings, the rendering of slow-motion feedback, the length of video scenes, and the function of scoring buttons.
- **Performance assessments:** Assessment of the own and the overall refereeing performance in real-life competitions.
- **Referee education with JudgED:** The contribution of JudgED in the context of referee education and its potential to complement practical referee training classes.
- **Target group of JudgED:** The potential of JudgED to enhance the skills of referees as well as its applicability to facilitate rule knowledge to athletes.
- **Application areas and usage of JudgED:** The potential application areas and usage modalities of JudgED.
- **General feedback:** General free text feedback of any kind to improve the quality of JudgED.

The particular questions of the questionnaire are listed below.

Willkommen zur Umfrage

Im Rahmen eines Projekts an der TU Wien, wird das Programm JudgED entwickelt. Ziel dieses Programms ist es, Entscheidungen von Schiedsrichtern in Kampfsportarten zu verbessern. In diesem Zusammenhang, würden wir gerne Ihre Meinungen bzw. Einschätzungen mithilfe eines Fragebogens erfahren.

Ihr Feedback hilft uns dabei, das erwähnte Programm zu verbessern und zu evaluieren. Vielen Dank für Ihre Teilnahme!

Obwohl Sie bereits die Gelegenheit hatten, JudgED im Rahmen der ÖSTM 2022 in Graz zu verwenden, sehen Sie sich bitte das folgende kurze Video über das Programm JudgED an:

Die Auswertung der Umfrage erfolgt streng vertraulich und Daten werden vor einer Veröffentlichung anonymisiert.

Basisinformationen

Erzählen Sie uns etwas über sich und Ihre Erfahrung als Schiedsrichter.

* 1. Was ist ihr Geburtsdatum?

Tag

Monat

Jahr

* 2. Was ist Ihr Geschlecht?

- ◯ Weiblich
- ◯ Männlich

* 3. In welchen Disziplinen besitzen Sie eine gültige Lizenz als Schiedsrichter?

- ☐ **Tatami** - Point fighting
- ☐ **Tatami** - Light contact
- ☐ **Tatami** - Kick light
- ☐ **Ring** - Full contact
- ☐ **Ring** - Low kick
- ☐ **Ring** - K1 Style

* 4. Was ist die höchste Kickboxing-Schiedsrichter-Lizenz, die Sie besitzen?

- ◯ **F** – National
- ◯ **E** – National
- ◯ **D** – National
- ◯ **C** – International
- ◯ **B** – International
- ◯ **A** – International
- ◯ **A1** – Internationaler Supervisor

* 5. In welchem Land haben Sie die letzte Kickboxing-Schiedsrichterausbildung absolviert?

- ◯ Österreich
- ◯ Deutschland
- ◯ Schweiz
- ◯ Andere

* 6. In welchem Jahr haben Sie die letzte Kickboxing-Schiedsrichterausbildung absolviert?

* 7. Wieviele Jahre haben Sie bereits Erfahrung als Kickboxing-Schiedsrichter?

0 Jahre 35 Jahre 70 Jahre

* 8. In wievielen Turniere pro Jahr nehmen Sie für gewöhnlich als Schiedsrichter teil?

- ◯ 1-3
- ◯ 4-7
- ◯ 8-12
- ◯ 13-15
- ◯ Mehr als 15

* 9. Seit wie vielen Jahren beschäftigen Sie sich bereits mit Kickboxen (als Athlet, Schiedsrichter oder Interessierter)?

0 Jahre 35 Jahre 70 Jahre

JudgED

Computerkenntnisse und Nutzungsgewohnheiten

Erzählen Sie uns etwas über Ihre Computerkenntnisse und Nutzungsgewohnheiten.

* 10. Wie würden Sie Ihre Computer-Fähigkeiten einschätzen?

- ◯ Sehr gut
- ◯ Gut
- ◯ Mittelmäßig
- ◯ Schlecht
- ◯ Sehr schlecht

* 11. Wie oft benutzen Sie einen Computer oder ein Tablet im Schnitt?

- ◯ Täglich
- ◯ Alle paar Tage
- ◯ Einmal pro Woche
- ◯ Seltener als einmal pro Woche

* 12. Sind Ihre Computerkenntnisse ausreichend, um ein Programm wie *JudgED* zu bedienen?

- ◯ Ja
- ◯ Nein

* 13. Haben Sie Zugang zu einem Gerät, um ein Programm wie *JudgED* zu bedienen (PC, Laptop oder Tablet)?

- ◯ Ja
- ◯ Nein

* 14. Haben Sie im Zuge Ihrer Schiedsrichterausbildung bereits ein Computerprogramm verwendet?

- ◯ Ja
- ◯ Nein

Nähere Information zu verwendeten Programmen

Sie haben angekreuzt, dass Sie im Zuge der Schiedsrichterausbildung bereits ein Computerprogramm verwendet haben. Bitte nennen Sie uns hierzu mehr Informationen.

* 15. Was war der Hauptzweck des Programms, welches im Zuge der Schiedsrichterausbildung verwendet wurde?

- ◯ File-Sharing-Plattform, um Lerninhalte herunterzuladen
- ◯ E-Learning-Plattform mit Folien und weiteren Funktionalitäten
- ◯ Regelerklärung- und Visualisierungs-Software
- ◯ Software zur Auswertung von echten Kampfsituationen
- ◯ Programm zur theoretischen Prüfung
- ◯ Programm zur praktischen Prüfung

16. Wenn Sie **2**, **4** oder **6** geantwortet haben – nennen Sie nähere Informationen zum verwendeten Programm (Name des Programms, Webseite, etc)?

Beitrag von Computersystemen zur Schiedsrichterausbildung

Erzählen Sie uns etwas über Ihre Einschätzungen bezüglich der Nutzung von digitalen Hilfsmitteln und Computersystemen im Zuge der Schiedsrichterausbildung.

* 17. Der Einsatz von digitalen Hilfsmitteln im Zuge der Schiedsrichterausbildung kann den Unterricht aufwerten.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu
- ◯ Keine der oben genannten

* 18. Der Einsatz von digitalen Hilfsmitteln im Zuge der Schiedsrichterausbildung kann den Lernerfolg verbessern.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 19. Die Verwendung eines Computersystems zur Schiedsrichterausbildung kann mir dabei helfen, meine Fähigkeiten hinsichtlich Regelkenntnis zu erweitern.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 20. Die Verwendung eines Computersystems zur Schiedsrichterausbildung kann mir dabei helfen, meine Fähigkeiten hinsichtlich praktischer Einschätzungen zu erweitern.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 21. Die Verwendung eines Computersystems zur Schiedsrichterausbildung kann meinen Kollegen dabei helfen deren Fähigkeiten hinsichtlich Regelkenntnis zu erweitern.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 22. Die Verwendung eines Computersystems zur Schiedsrichterausbildung kann meinen Kollegen dabei helfen deren Fähigkeiten hinsichtlich praktischer Einschätzungen zu erweitern.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

Wettbewerb im Zuge der Schiedsrichterausbildung

Teilen Sie uns mit, welchen Einfluss der Einsatz von Leistungs-Vergleichen und Ranglisten im Zuge Ihrer Schiedsrichterausbildung hinsichtlich Motivation und Spaß gehabt hätte.

* 23. Ein anonymer Vergleich meiner erlernten Schiedsrichter-Fähigkeiten mit anderen Kollegen hätte meine Lern-Motivation während der Schiedsrichterausbildung erhöht.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 24. Ein anonymer Vergleich meiner erlernten Schiedsrichter-Fähigkeiten mit anderen Kollegen hätte den Spaß während der Schiedsrichterausbildung erhöht.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 25. Eine personalisierte Rangliste unter Schiedsrichter-Kollegen hätte meine Lern-Motivation beim Lernen während der Schiedsrichterausbildung erhöht.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 26. Eine personalisierte Rangliste unter Schiedsrichter-Kollegen hätte den Spaß während der Schiedsrichterausbildung erhöht.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

Messwerte in *JudgED*

In *JudgED* werden insbesondere Reaktionszeit und Entscheidungsgenauigkeit angezeigt, um Statistiken zu berechnen. Geben Sie uns einen Einblick in Ihre Ansichten diesbezüglich.

* 27. Als Schiedsrichter ist es wichtig, schnelle Entscheidungen während den Kämpfen zu treffen.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu
- ◯ Nur in bestimmten Disziplinen

* 28. Als Schiedsrichter ist es wichtig, Situationen nicht nur isoliert zu betrachten, sondern auch vorhergehende Situationen in Entscheidungen zu berücksichtigen (Kontext).

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 29. **Reaktionszeit:** Zeitdifferenz zwischen dem Auftreten eines zu bewertenden Treffers und der Bewertung durch den Schiedsrichter.

Die Anzeige von Statistiken hinsichtlich Reaktionszeit in *JudgED* ist sinnvoll.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 30. **Entscheidungsgenauigkeit:** Verhältnis von richtigen und falschen Schiedsrichter-Entscheidungen.

Die Anzeige von Statistiken hinsichtlich Entscheidungsgenauigkeit in *JudgED* ist sinnvoll.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 31. **Vergleiche:** Vergleich der eigenen Leistungen (Reaktionszeit, Entscheidungsgenauigkeit) mit anderen Schiedsrichtern.

Der Vergleich mit Statistiken anderer *JudgED*-Benutzer ist sinnvoll.

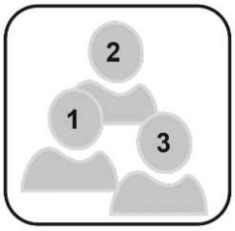
- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

Varianten von Features in *JudgED*

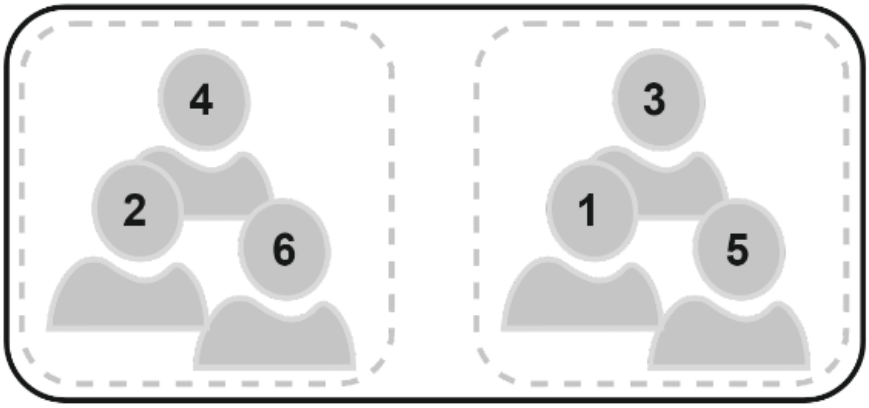
Einige Features in *JudgED* können zukünftig auf verschiedene Weisen umgesetzt werden. Teilen Sie uns Ihre Meinung zu ausgewählten Funktionen in *JudgED* mit.

* 32. Auf welchen Personenkreis sollen sich Ranglisten in *JudgED* beziehen?
Sollten verschiedene Ranglisten sinnvoll sein, wählen Sie bitte mehrere aus.

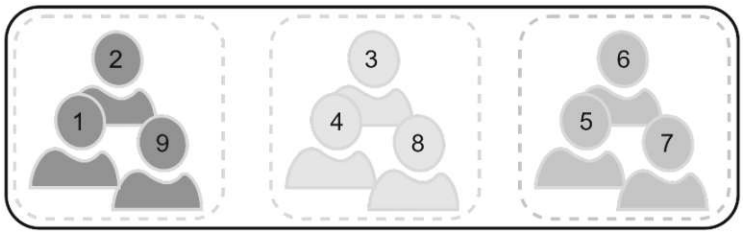
☐ **Teilnehmer eines Kurses:**
Kollegen der laufenden Ausbildung



☐ **Benutzer der gleichen Disziplin:**
Alle Teilnehmer von vergangenen und laufenden Kursen der gleichen Disziplin



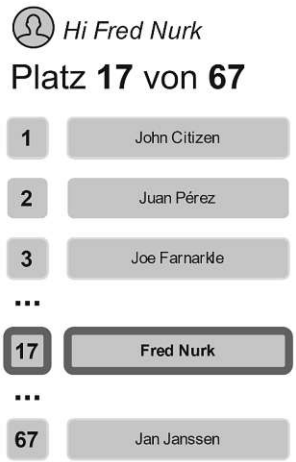
☐ **Benutzer aller Disziplinen:**
Alle Teilnehmer von vergangenen und laufenden Kursen über alle Disziplinen



☐ **Keine Rangliste:**
Eine Rangliste ist nicht notwendig

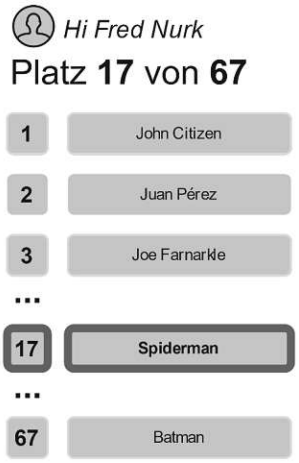
* 33. Wie sollen Ranglisten hinsichtlich der Privatsphäre in *JudgED* gestaltet werden?

○ **Vollständig personalisierte Rangliste:**
Vorname + Nachname + Rang ist in der Rangliste ersichtlich



○ **Personalisierte Rangliste der Top-Plätze – Variante 1:**
Vorname + Nachname + Rang der Top-Plätze ist in der Rangliste ersichtlich
(andere Ränge nur privat sichtbar)

○ **Personalisierte Rangliste der Top-Plätze – Variante 2:**
Vorname + Nachname + Rang der Top-Plätze ist in der Rangliste ersichtlich
(andere Ränge nur als Pseudonym sichtbar)



○ **Vollständig pseudonymisierte Rangliste:**
Pseudonym + Rang ist in der Rangliste ersichtlich

○ **Pseudonymisierte Rangliste der Top-Plätze:**
Pseudonym + Rang der Top-Plätze ist in der Rangliste ersichtlich
(andere Ränge nur privat sichtbar)

○ **Keine Rangliste:**
Eigenen Rang nur privat anzeigen

* 34. Um die Nachvollziehbarkeit einer Punkte-Entscheidungen zu ermöglichen, wird in *JudgED* eine Zeitlupe der Szene angezeigt.

Wie würden Sie die Zeitlupen-Geschwindigkeit der folgenden Szene einschätzen?

- ◯ Zu langsam
- ◯ Zu schnell
- ◯ Genau richtig

* 35. Um die Nachvollziehbarkeit einer Punkte-Entscheidungen zu ermöglichen, wird in *JudgED* eine Zeitlupe der Szene angezeigt.

Wie würden Sie die Zeitlupen-Länge der folgenden Szene einschätzen?

- ◯ Zu kurz
- ◯ Zu lange
- ◯ Genau richtig

* 36. Die Hauptaufgabe von *JudgED*-Benutzern besteht darin, eine Reihe an Video-Szenen so schnell und genau wie möglich zu bewerten. Für die Disziplin *Point fighting* ist eine Länge von ca. 4-15 Sekunden für die einzelnen Videos geplant.

Wie würden Sie die Länge dieser Videos einschätzen?

- ◯ Zu lang
- ◯ Zu kurz
- ◯ Genau richtig
- ◯ Ich weiß nicht (ich bewerte kein Point fighting)

* 37. Die Hauptaufgabe von *JudgED*-Benutzern besteht darin, eine Reihe an Video-Szenen so schnell und genau wie möglich zu bewerten. Für *Running-Time-Disziplinen* ist eine Länge von 45-80 Sekunden für die einzelnen Videos geplant.

Wie würden Sie die Länge dieser Videos einschätzen?

- ◯ Zu lang
- ◯ Zu kurz
- ◯ Genau richtig
- ◯ Ich weiß nicht (ich bewerte kein Running-Time)

* 38. Je nach Disziplin, stehen dem Benutzer in *JudgED* verschiedene Tasten zur Verfügung, um Punkte für einen der beiden Kämpfer zu vergeben. Zusätzlich gibt es auch die Möglichkeit eine Strafe (Penalty) zu vergeben. Hierbei ist geplant, dass es nur eine *Verwarnung-Taste* pro Athlet gibt, die für „*Warning*“ und „*Exit*“ steht. Dieser Ansatz wurde vorgeschlagen, um die Anzahl der Tasten zu verringern.

Was halten Sie von diesem Ansatz?

- ◯ Warning und Exit sollten als separate Tasten existieren
- ◯ Weder Warning, noch Exit sollte auswählbar sein, da es sonst zu kompliziert ist
- ◯ Ja, eine Taste für beides sollte genügen

Selbsteinschätzung der Schiedsrichter-Kompetenzen in Wettbewerben

Teilen Sie uns mit, wie Sie Ihre Leistung als Schiedsrichter in Wettbewerben einschätzen.

* 39. Wie schätzen Sie Ihre Leistung als Schiedsrichter im Allgemeinen ein?

- ◯ Sehr gut
- ◯ Gut
- ◯ Mittelmäßig
- ◯ Schlecht
- ◯ Sehr schlecht

* 40. Wie schätzen Sie Ihre Leistung als Schiedsrichter im Vergleich zu anderen Schiedsrichtern ein?

- ◯ Überdurchschnittlich
- ◯ Durchschnittlich
- ◯ Unterdurchschnittlich

* 41. Ich bin überzeugt von meinen Fähigkeiten als Schiedsrichter.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 42. Ich bin in der Lage, akkurate Entscheidungen als Schiedsrichter zu treffen.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 43. Ich fühle mich der Herausforderung gewachsen, die an mich gestellen Anforderungen als Schiedsrichter zu erfüllen.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 44. **Reaktionszeit:** Zeitdifferenz zwischen dem Auftreten eines zu bewertenden Treffers und der Bewertung durch den Schiedsrichter.

Wie würden Sie die durchschnittliche Reaktionszeit aller Schiedsrichter in Wettkämpfen Ihrer Disziplinen einschätzen?

- ◯ 0,2 - 0,5 Sekunden
- ◯ 0,5 - 0,8 Sekunden
- ◯ 0,8 - 1,1 Sekunden
- ◯ 1,1 - 1,4 Sekunden
- ◯ 1,4 - 1,7 Sekunden
- ◯ 1,7 - 2,0 Sekunden
- ◯ 2,0 - 2,3 Sekunden
- ◯ 2,3 - 2,6 Sekunden
- ◯ 2,6 - 3,0 Sekunden
- ◯ > 3,0 Sekunden

* 45. **Reaktionszeit:** Zeitdifferenz zwischen dem Auftreten eines zu bewertenden Treffers und der Bewertung durch den Schiedsrichter.

Wie würden Sie Ihre durchschnittliche Reaktionszeit in Wettkämpfen einschätzen?

- ◯ 0,2 - 0,5 Sekunden
- ◯ 0,5 - 0,8 Sekunden
- ◯ 0,8 - 1,1 Sekunden
- ◯ 1,1 - 1,4 Sekunden
- ◯ 1,4 - 1,7 Sekunden
- ◯ 1,7 - 2,0 Sekunden
- ◯ 2,0 - 2,3 Sekunden
- ◯ 2,3 - 2,6 Sekunden
- ◯ 2,6 - 3,0 Sekunden
- ◯ > 3,0 Sekunden

* 46. **Entscheidungsgenauigkeit:** Verhältnis von richtigen und falschen Schiedsrichter-Entscheidungen.

Wie würden Sie die durchschnittliche Entscheidungsgenauigkeit aller Schiedsrichter in Wettkämpfen Ihrer Disziplinen einschätzen?

- ◯ Kleiner als 50 %
- ◯ 50 - 60 %
- ◯ 60 - 70 %
- ◯ 70 - 80 %
- ◯ 80 - 90 %
- ◯ 90 - 95 %
- ◯ 95 - 100 %

* 47. **Entscheidungsgenauigkeit:** Verhältnis von richtigen und falschen Schiedsrichter-Entscheidungen.

Wie würden Sie Ihre durchschnittliche Entscheidungsgenauigkeit in Wettkämpfen einschätzen?

- ◯ Kleiner als 50 %
- ◯ 50 - 60 %
- ◯ 60 - 70 %
- ◯ 70 - 80 %
- ◯ 80 - 90 %
- ◯ 90 - 95 %
- ◯ 95 - 100 %

JudgED - Einsatz im Zuge der Schiedsrichterausbildung

Teilen Sie uns Ihre Ansichten über den potenziellen Beitrag von *JudgED* zur Schiedsrichterausbildung mit.

* 48. Der Einsatz von *JudgED* im Zuge der Schiedsrichterausbildung kann den Unterricht aufwerten.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 49. Der Einsatz von *JudgED* im Zuge der Schiedsrichterausbildung kann den Lernerfolg verbessern.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 50. Der Einsatz von *JudgED* kann den praxisbezogenen Teil der Schiedsrichterausbildung vorort ergänzen.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 51. Der Einsatz von *JudgED* kann den praxisbezogenen Teil der Schiedsrichterausbildung vorort ersetzen.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

JudgED

Zielgruppe von JudgED

Einschätzungen für welche Personenkreise JudgED eingesetzt werden kann.

* 52. Der Einsatz von *JudgED* kann für unerfahrene Schiedsrichter sinnvoll sein, um deren Fähigkeiten zu erweitern.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 53. Der Einsatz von *JudgED* kann für erfahrene Schiedsrichter sinnvoll sein, um deren Fähigkeiten zu erweitern.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 54. Der Einsatz von *JudgED* kann für Athleten sinnvoll, um deren Regelverständnis zu verbessern.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

Zukünftige Anwendungsgebiete und Nutzung *JudgED*

Einschätzungen in welcher Form *JudgED* voraussichtlich genutzt werden kann.

* 55. In welchem Modus sehen Sie die zukünfige Anwendung von *JudgED*?

- ☐ **Während Kursen/Seminaren:** Für praktische Übungen zwischendurch
- ☐ **Nach Kursen/Seminaren:** Als zusätzliche Übungsmethode
- ☐ **Unabhängig von Kursen:** Zum selbstständigen Auffrischen der Fähigkeiten
- ☐ Sonstiges (bitte angeben)

* 56. Auf welchem Gerät würden Sie *JudgED* am ehesten benutzen?

- ◯ Laptop/PC
- ◯ Tablet
- ◯ Smartphone

* 57. Wie oft würden Sie *JudgED* freiwillig verwenden?

- ◯ Täglich
- ◯ Alle paar Tage
- ◯ Einmal pro Woche
- ◯ Seltener als einmal pro Woche
- ◯ Gar nicht

Feedback

Im Zuge der ÖSTM 2022 in Graz hatten Sie bereits die Gelegenheit *JudgED* in einer eingeschränkten Form zu testen. Wir würden Sie bitten, Ihre Erfahrungen mit uns zu teilen, sodass wir gezielte Verbesserungen vorzunehmen können.

58. Bitte nennen und beschreiben Sie alle Punkte, welche an *JudgED* verbessert werden sollen (z.B. fehlende Funktionalitäten, verbesserungswürdige Funktionalitäten, Benutzerfreundlichkeit, etc).

Questionnaire 2

This questionnaire intends to assess motivation, experience and feelings associated with playing the serious game. While the questionnaire in Appendix 7 was conducted after the case study, this questionnaire was not applied in the course of this work, as the participants in the case study were only exposed to a subset of the serious game's functionalities. The conduction of this questionnaire presumes the extensive usage of the serious game's functionalities over a significant period. Therefore, it is only applicable as a post-experimental instrument in long-term field experiments.

The questionnaire comprises questions from the standard *Post-Experimental Intrinsic Motivation Inventory scale* (IMI) [inta]. It was built by selecting, customizing and translating already phrased standard questions. From the complete set of seven subscales and 45 question items included in the IMI scale, 24 items from six subscales are used in this questionnaire:

- **Interest/Enjoyment:** The intrinsic motivation for using JudgED measured by the level of fun, enjoyment, and interest felt while playing the serious game.
- **Perceived competence:** The level of satisfaction and progress induced by practising decision-making skills with JudgED and the perceived performance compared to other players.
- **Effort/Importance:** The level of effort invested in the training with JudgED and the importance of doing well in the requested tasks.
- **Pressure/Tension:** The degree of perceived pressure while judging the video scenes and the existence of self-doubt to perform well in the requested tasks.
- **Value/Usefulness:** The willingness to regularly use JudgED, its applicability to training decision-making skills, and the ability to transfer acquired skills to real-life contests.
- **Perceived Choice:** The degree to which the usage of JudgED was perceived voluntary.

While the questions included in the following questionnaire are sorted by the above categories, they should be arranged in random order before conducting the survey [inta].

Willkommen zur Umfrage

In den letzten drei Wochen hatten Sie die Gelegenheit, sich näher mit JudgED zu beschäftigen. Berichten Sie uns mehr über Ihre Erfahrungen und Einschätzungen.

Ihr Feedback hilft uns dabei, das erwähnte Programm zu verbessern und zu evaluieren. Vielen Dank für Ihre Teilnahme!

Die Auswertung der Umfrage erfolgt streng vertraulich und Daten werden vor einer Veröffentlichung anonymisiert.

* 1. Das Training mit *JudgED* hat mir Spaß gemacht.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 2. Das Training mit *JudgED* war sehr interessant.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 3. Das Training mit *JudgED* war sehr unterhaltsam.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 4. Ich bin zufrieden mit meinen Leistungen in *JudgED*.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zufrieden

* 5. Ich glaube, ich war beim Bewerten der Video-Szenen in *JudgED* ziemlich gut.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 6. Ich glaube, verglichen mit anderen, war ich beim Bewerten der Video-Szenen in *JudgED* ziemlich gut.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 7. Nach einiger Zeit der Übung, entwickelte ich eine Sicherheit in der Ausübung der geforderten Tätigkeiten in *JudgED*.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 8. Beim Bewerten der Video-Szenen in *JudgED* stelle ich mich geschickt an.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 9. Ich habe viel Zeit in das Training mit *JudgED* investiert.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 10. Ich habe hart daran gearbeitet, beim Bewerten der Video-Szenen in *JudgED* gut abzuschneiden.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 11. Es war mir wichtig, beim Bewerten der Video-Szenen in *JudgED* gut abzuschneiden.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 12. Beim Bewerten der Video-Szenen in *JudgED* fühlte ich mich unter Druck gesetzt.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 13. Beim Bewerten der Video-Szenen in *JudgED* fühlte ich mich angespannt.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 14. Ich hatte Bedenken, ob ich die geforderten Tätigkeiten in *JudgED* gut hinbekomme.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 15. Die Verwendung von *JudgED* könnte für mich nützlich sein.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 16. Der Einsatz von *JudgED* könnte nützlich sein, um die Entscheidungsgenauigkeit von Judges zu verbessern.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 17. Der Einsatz von *JudgED* könnte nützlich sein, um die Reaktionszeit von Judges zu verbessern.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 18. Ich wäre bereit *JudgED* regelmäßig zu nutzen, da es für mich einen gewissen Wert hat.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 19. Das Trainieren mit *JudgED* könnte mir dabei helfen, meine Entscheidungen in Wettkämpfen zu verbessern.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 20. Das Trainieren mit *JudgED* könnte mir dabei helfen, meine Reaktionszeit in Wettkämpfen zu verbessern.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 21. Ich glaube, der Einsatz von *JudgED* ist wichtig.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 22. Ich hatte das Gefühl, dass das Training mit *JudgED* freiwillig zu machen.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 23. Ich habe *JudgED* verwendet, weil ich es wollte.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

* 24. Ich habe *JudgED* verwendet, weil ich es musste.

- ◯ Stimme völlig zu
- ◯ Stimme zu
- ◯ Stimme weder zu noch nicht zu
- ◯ Stimme nicht zu
- ◯ Stimme überhaupt nicht zu

JSON Schema of Entities

JSON Schema - Video


```
{
  "$schema": "https://json-schema.org/draft/2019-09/schema",
  "$id": "https://reg-web-prod.herokuapp.com/video.json",
  "title": "Record of video",
  "type": "object",
  "properties": {
    "key": {
      "title": "Unique identifier for a video",
      "type": "string"
    },
    "basics": {
      "title": "Characteristics of a video",
      "type": "object",
      "properties": {
        "association": {
          "title": "Association of the video",
          "type": "string",
          "enum": [
            "wako",
            "wkf"
          ]
        },
        "discipline": {
          "title": "Discipline of the video",
          "type": "string",
          "enum": [
            "pointfighting",
            "light_contact",
            "kick_light",
            "full_contact",
            "full_contact_low_kick",
```

```
        "full_contact_k1",
        "kumite"
      ]
    },
    "tournament": {
      "title": "Tournament depicted in the video",
      "type": "string"
    },
    "year": {
      "title": "Year of the tournament",
      "type": "number"
    },
    "ageGroup": {
      "title": "Age group of the fight",
      "type": "string",
      "enum": [
        "v",
        "s",
        "j",
        "oc",
        "yc"
      ]
    },
    "surnameAthlete1": {
      "title": "Name of the red athlete",
      "type": "string"
    },
    "surnameAthlete2": {
      "title": "Name of the blue athlete",
      "type": "string"
    },
    "gender": {
      "title": "Gender of the athletes",
      "enum": [
        "m",
        "f"
      ]
    },
    "bout": {
      "title": "The bout type of the video",
      "type": "string",
      "enum": [
        "round_1",
```

```
            "round_2",
            "round_3",
            "full_fight"
          ]
        }
      }
    },
    "vimeo": {
      "title": "Information about the video file stored in Vimeo",
      "type": "object",
      "properties": {
        "accessLink": {
          "title": "URL to access the video in the browser",
          "type": "string"
        },
        "playerEmbedUrl": {
          "title": "URL to access the video via the Vimeo JS player",
          "type": "string"
        },
        "resourceKey": {
          "title": "The Vimeo resource key of the video",
          "type": "string"
        },
        "uri": {
          "title": "The Vimeo URI suffix in the format /videos/id",
          "type": "string"
        },
        "id": {
          "title": "The Vimeo video ID",
          "type": "string"
        },
        "transcoded": {
          "title": "Whether or not Vimeo has transcoded the video",
          "type": "boolean"
        },
        "duration": {
          "title": "Duration of the video",
          "type": "number"
        }
      }
    },
    "metadata": {
      "title": "Audit metadata about the object",
```

```
      "type": "object"
    },
    "thumbnails": {
      "title": "Thumbnail of the video in various sizes",
      "type": "object"
    },
    "firstScoreThumbnails": {
      "title": "Thumbnail of first decision in video",
      "type": "object"
    }
  }
}
```

JSON Schema - Video Scene

```
{
  "$schema": "https://json-schema.org/draft/2019-09/schema",
  "$id": "https://reg-web-prod.herokuapp.com/videoScene.json",
  "title": "Record of video scene",
  "type": "object",
  "properties": {
    "key": {
      "title": "Unique identifier for a video scene",
      "type": "string"
    },
    "basics": {
      "title": "Characteristics of a video scene",
      "type": "object",
      "properties": {
        "sceneStart": {
          "title": "Start time in context of linked video",
          "type": "number"
        },
        "sceneEnd": {
          "title": "End time in context of linked video",
          "type": "number"
        },
        "difficulty": {
          "title": "Estimated difficulty of the video scene",
          "type": "string",
          "enum": [
            "veryLow",
            "low",
```

```
        "medium",
        "high",
        "veryHigh"
      ]
    },
    "mainFocus": {
      "title": "The main focus of the video scene",
      "type": "array",
      "items": {
        "title": "Main focus tag",
        "type": "string"
      }
    },
    "trailingPeriod": {
      "title": "Seconds the video scene is extended",
      "type": "number"
    },
    "athleteLeftPosition": {
      "title": "Color of left-positioned athlete at scene start",
      "type": "string",
      "enum": [
        "blue",
        "red"
      ]
    },
    "blurring": {
      "title": "Blurring definition to cover referee guestures",
      "type": "object",
      "properties": {
        "visibilityStart": {
          "title": "Start time of showing blurring rectangles",
          "type": "number"
        },
        "visibilityEnd": {
          "title": "End time of showing blurring rectangles",
          "type": "number"
        },
        "drawing": {
          "title": "Blurring rectangles definition",
          "type": "array",
          "items": {
            "title": "One blurring rectangle",
            "type": "object",
```

```
                "properties": {
                  "left": {
                    "title": "X position rel. to video frame",
                    "type": "number"
                  },
                  "top": {
                    "title": "Y position rel. to video frame",
                    "type": "number"
                  },
                  "width": {
                    "title": "Width of the rectangle",
                    "type": "number"
                  },
                  "height": {
                    "title": "Height of the rectangle",
                    "type": "number"
                  }
                }
              }
            }
          }
        }
      },
      "scoreDefinitions": {
        "title": "Defined decisions at certain points in time",
        "type": "array",
        "items": {
          "title": "Defined decision",
          "type": "object",
          "properties": {
            "time": {
              "title": "Time of the defined decision",
              "type": "number"
            },
            "athlete": {
              "title": "Athlete to whom decision is assigned",
              "type": "string",
              "enum": [
                "blue",
                "red"
              ]
            },
```

```
"score": {
  "title": "Value of the decision (score or penalty)",
  "type": "string",
  "enum": [
    "0",
    "1",
    "2",
    "3",
    "P",
    "C1",
    "C2"
  ]
},
"technique": {
  "title": "Technique causing the decision",
  "type": "string"
},
"techniqueFreeText": {
  "title": "Technique causing the decision - free text",
  "type": "string"
},
"highlighting": {
  "title": "Highlighting definition for crucial areas",
  "type": "object",
  "properties": {
    "visibilityStart": {
      "title": "Start time of showing the highlighting",
      "type": "number"
    },
    "visibilityEnd": {
      "title": "End time of showing the highlighting",
      "type": "number"
    },
    "drawing": {
      "title": "The list of highlighting ellipsis",
      "type": "array",
      "items": {
        "title": "One highlighting definition",
        "type": "object",
        "properties": {
          "left": {
            "title": "X position rel. to video frame",
            "type": "number"
```

```
                    },
                    "top": {
                      "title": "Y position rel. to video frame",
                      "type": "number"
                    },
                    "width": {
                      "title": "Width of the ellipse",
                      "type": "number"
                    },
                    "height": {
                      "title": "Height of the ellipse",
                      "type": "number"
                    }
                  }
                }
              }
            }
          },
          "thumbnails": {
            "title": "Thumbnail at defined definition time",
            "type": "object"
          },
          "concurrent": {
            "title": "Whether decision is concurrent",
            "type": "boolean"
          }
        }
      }
    },
    "video": {
      "title": "Referenced video belonging to video scene",
      "type": "string"
    },
    "statusInfo": {
      "title": "Object to track changes of the video scene",
      "type": "object",
      "properties": {
        "status": {
          "title": "The current status",
          "type": "string"
        },
        "comment": {
          "title": "The current comment",
```

```
            "type": "string"
          },
          "history": {
            "title": "List of all current and history changes",
            "type": "array",
            "items": {
              "title": "Change history object",
              "type": "object",
              "properties": {
                "fromStatus": {
                  "title": "Source status",
                  "type": "string"
                },
                "toStatus": {
                  "title": "Target status",
                  "type": "string"
                },
                "comment": {
                  "title": "Comment of status change",
                  "type": "string"
                },
                "uid": {
                  "title": "UID of the user performing the change",
                  "type": "string"
                },
                "email": {
                  "title": "Email of the user performing the change",
                  "type": "string"
                },
                "date": {
                  "title": "Date when the change was performed",
                  "type": "data-time"
                }
              }
            }
          }
        },
        "metadata": {
          "title": "Audit metadata about the object",
          "type": "object"
        }
      }
```

```
}
```

JSON Schema - Playlist

```
{
  "$schema": "https://json-schema.org/draft/2019-09/schema",
  "$id": "https://reg-web-prod.herokuapp.com/playlist.json",
  "title": "Record of playlist",
  "type": "object",
  "properties": {
    "key": {
      "title": "Unique identifier for a playlist",
      "type": "string"
    },
    "association": {
      "title": "Association of included video scenes",
      "type": "string",
      "enum": [
        "wako",
        "wkf"
      ]
    },
    "name": {
      "title": "The name of the playlist",
      "type": "string"
    },
    "scoredScenesKeys": {
      "title": "The list of included video scene keys",
      "type": "array",
      "items": {
        "title": "Video scene key",
        "type": "string"
      }
    },
    "feedbackMode": {
      "title": "The feedback mode of the playlist",
      "type": "string",
      "enum": [
        "standard",
        "lab",
        "exam"
      ]
    },
```

```
    "metadata": {
      "title": "Audit metadata about the object",
      "type": "object"
    }
  }
}
```

JSON Schema - Course

```
{
  "$schema": "https://json-schema.org/draft/2019-09/schema",
  "$id": "https://reg-web-prod.herokuapp.com/course.json",
  "title": "Root Schema",
  "type": "object",
  "properties": {
    "key": {
      "title": "Unique identifier for a course",
      "type": "string"
    },
    "association": {
      "title": "The association of the course",
      "type": "string",
      "enum": [
        "wako",
        "wkf"
      ]
    },
    "disciplines": {
      "title": "The disciplines involved in the course",
      "type": "array",
      "items": {
        "title": "A Schema",
        "type": "string",
        "enum": [
          "pointfighting",
          "light_contact",
          "kick_light",
          "full_contact",
          "full_contact_low_kick",
          "full_contact_k1",
          "kumite"
        ]
      }
```

```
    },
    "fromDate": {
      "title": "Start date of the course",
      "type": "date-time"
    },
    "toDate": {
      "title": "End date of the course",
      "type": "date-time"
    },
    "name": {
      "title": "Name of the course",
      "type": "string"
    },
    "open": {
      "title": "Whether course is open for all users",
      "type": "boolean"
    },
    "participants": {
      "title": "The participants (users) of the course",
      "type": "array",
      "items": {
        "title": "UID of the course participant (user)",
        "type": "string"
      }
    },
    "playlists": {
      "title": "The playlists assigned to this course",
      "type": "array",
      "items": {
        "title": "Key of the playlist",
        "type": "string"
      }
    },
    "metadata": {
      "title": "Audit metadata about the object",
      "type": "object"
    }
  }
}
```

JSON Schema - User

```
{
  "$schema": "https://json-schema.org/draft/2019-09/schema",
  "$id": "https://reg-web-prod.herokuapp.com/user.json",
  "title": "Record of user",
  "type": "object",
  "properties": {
    "uid": {
      "title": "Firebase UID of the user",
      "type": "string"
    },
    "firstName": {
      "title": "First name of the user",
      "type": "string"
    },
    "secondName": {
      "title": "Second name of the user",
      "type": "string"
    },
    "metadata": {
      "title": "Audit metadata about the object",
      "type": "object"
    }
  }
}
```