FACULTY OF TECHNOLOGY

Benchmarking framework for serious games in project management

Laura Kauppinen

Industrial engineering and management

Master's thesis

April 2022

ABSTRACT

Benchmarking framework for serious games in project management

Laura Kauppinen

University of Oulu, Industrial engineering and management

Master's thesis 2022, 76 pp. + 2 appendixes

Supervisor(s) at the university: Jere Lehtinen

Serious games are games that are developed with educational goals instead of entertainment at the core. There have been several different ways serious games have been assessed, but often these targets a specific part of the game. This thesis aims to construct a benchmarking framework for evaluating serious games. The evaluated games come from the project management domain, and as such, the framework is directed to said area. The evaluations of the games are done through action research orientation, with benchmarking as the method to evaluate games and the framework itself. The construction of the framework is based on an exploratory literature review. The gathered framework consists of the following evaluation areas: game session, mechanics, playability and flow, project management and costs. The evaluation of serious games can not be based on the game elements alone. However, consideration must be given to the materials related to the game and the game's fit to the learning event. Often the marketing for serious games was focused on the subject materials of the game, while the mechanics and playability of the game were left to lower attention. More research is needed on the usability of the framework and the evaluations on the current state of project management serious games. The developed framework's core could also be tested against different serious game domains.

Keywords: Benchmarking, serious games, project management

Abstrakti

Projektijohtamisen vakavien pelien benchmarkkaus arviointimalli

Laura Kauppinen

Oulun Yliopisto, Tuotantotalous

Diplomityö 2022, 76 ss. + 2 appendixes

Supervisor(s) at the university: Jere Lehtinen

Vakavia pelejä kehittäessä oppimistavoitteet ovat keskiössä viihdearvon sijasta. Ajan saatossa on muotoiltu lukuisia eri tapoja arvioida vakavia pelejä, mutta usein nämä arviot kohdistuvat yksittäiseen osa-alueeseen pelissä. Tämän diplomityön tavoitteena on muodostaa benchmarkkaus arviointimalli vakavien pelien arvioimiseksi. Arvioitavat pelit kerätään projektijohtamisen alueelta. Tämän vuoksi arviointimallin opetuspelien aiheeseen liittyvät osa-alueet kohdistuvat projektijohtamisen osa-alueelle. Arvioinneissa hyödynnetään toimintatutkimuksen menetelmiä täydentämään benchmarkkauksen toimintatapoja. Tutkimuksessa arviointi kohdistuu niin peleihin kuin malliin itseensä. Pelien benchmarkkaus arviointimalli koostettiin tutkivan kirjallisuuskatsauksen pohjilta. Arviointimalli kohdistuu viidestä osa-alueesta: pelitapahtuma, pelin mekaniikka, pelattavuus ja pelin luontevuus, projektijohtaminen ja kustannukset. Pelien arviointi ei voi pohjata pelkästään pelin osa-alueisiin, vaan on välttämätöntä tarkastella myös pelin soveltuvuutta pelitapahtumaan ja käytettävissä oleviin resursseihin. Usein vakavien pelien markkinointi keskittyy pelin opetettavan aiheen ympärille, jättäen pelin mekaniikan ja pelattavuuden toissijaisiksi. Kehitettyä benchmarkkaus arviointimallia tulisi arvioida lisää jatkotutkimuksissa. Lisäksi arviointimallia voisi soveltaa myös muilla aloilla, kuin projektijohtamisessa.

Avainsanat: Benchmarkkaus, vakavat pelit, projektijohtaminen

FOREWORD

This thesis aims to develop a benchmarking framework for the evaluation of serious games. My interest in the topic arose from my interest in the usage of video games in different domains outside just entertainment value. The thesis work spanned from October of 2021 to start of April 2022.

I would like to thank my supervisor from the university side, Jere Lehtinen, for the discussion on the topic and guidance in the thesis process. The comments provided helped to shape the thesis into a more logical form and to formulate the core on which the thesis is constructed. I also thank my friends for aiding in shaping the eventual framework. And finally, I would like to thank my partner for proofreading, acting as a debating partner and supporting me through this process.

Oulu, 04.04.2022

Laura Kauppinen
Author

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

BoK	Boby of knowledge
Co	Costs
GS	Game Session
Me	Mechanics
PF	Playability and flow
PM	Project management
PMI	Project management institute
PRINCE2	Projects in controlled environment 2
SBL	Simulation based learning

1 INTRODUCTION

Serious games can be defined in several different ways. Lim *et al.* (2013) defined serious games as game artefacts developed for learning purposes with an impact on the learners through other methods. Almeida and Simoes (2019) noted that serious games have goal-oriented tasks that focus on improving players' skills or performance in some scenarios. The focal area of serious games design is the educational aspect of the game instead of entertainment value (M.Nazry and Romano 2017). Educational games are games that serve an educational purpose. The original design goal of the game might not be education, but it fits this purpose in its' use (Noemí and Máximo 2014).

A related term to serious games is gamification. Gamification takes game elements, such as leader boards, experience point accumulation, or badges, and applies them in a non-gaming context, such as a course environment, to change the learning experience (Lameras et al. 2017). Simulation-based learning (SBL) is one of the other adjacent terms related to serious games. SBL emerged from the business games as educators transformed them into simulations by increasing lifelike features of the game to shape them more towards work context (Hallinger and Wang 2020a).

These closely related terms are often mixed (Hallinger and Wang 2020a). Lameras *et al.* (2017) defined gamification to be part of the serious games umbrella, while Lim *et al.*(2013), M.Nazry and Romano(2017), and Almeida and Simoes (2019) draw the line between usage of game and game elements as the line between serious games and gamification. Using a game instead of just game elements is part of the differentiation between gamification and serious games. Game components and their interaction provide a basis for analysing the serious game from a gameplay perspective.

The field of serious games is broad and is not limited to only video games but also includes tabletop board games along with role-playing, simulation environments and paper-based notebook games. As such, it is essential to evaluate the fit of the serious game on a case by case basis (Girard et al. 2013).

Hallinger and Wang (2020a) noted that most active simulation-based learning authors work in either medical or management contexts. Some papers, though, cover more general topics such as academic writing, but as the research carried out by O'Flaherty and Costabile (2020) was done in nursing school, this would align it with the medical side. Hallinger and Wang (2020a) also noted that medical and management fields often progress on the same tracks without interaction with one another. The lack of interaction resulted in similar topics and mistakes repeated, while the other track developed solutions and best practices for the error and challenges earlier. As benchmarking has been used to extend to other parties in the industry, including non-competitors, the development of benchmarking model connecting gameplay and serious games elements, could provide a tool to assess games regardless of their topic and compare the best practices (Zairi and Leonard 1994). As Hallinger and Wang (2020a) mentioned, sticking to academic branches has been a noted as a weakness of simulation based learning and serious games literature. Here the benchmarking methods could provide valuable tool to reassess the information from different areas of academia.

1.1 Research problem, objectives and scope

The hypothesis for this thesis is that for evaluation of serious games through benchmarking practice can be based on the game components, such as mechanics. The benchmarking practice helps to find characteristics of a good serious game and enables evaluation and finding of the best practices for a serious game in project management education. This hypothesis is examined based on two research questions:

1. What kind of features of serious games can be used to benchmark from a game component perspective?
2. How do different project management games compare to these features?

The first question's answer formulates the base for the benchmarking model used to evaluate the games and simulations. The second question provides the method of verifying the proposed framework and assessing its weaknesses. The evaluated games are selected from the project management field, focusing on simulation games. The benchmarking process is carried around the serious games in project management, but

the goal is to apply the framework in other serious games or simulation evaluations. The focus area selection is just made to ease comparisons between the evaluated games. In this thesis, the games being assessed are all videogames to facilitate the comparison and reduce variables, but the model is not limited to only videogames at its core.

This thesis aims to formulate a benchmarking model for serious games. The selected scope of the games in the thesis is project management games. The broad subject matter provides plenty of games to benchmark with clear points for comparison. Selection of point of view also provides more straightforward evaluation options of the model. For example, are some of the aspects weighted too much or little. When formulating the benchmarking framework, the goal is to keep it as open as possible, with limited parts targeting the topic the game is teaching. This way, the framework can be applied to several different types of serious games with minor modifications based on the expertise of the teaching staff.

In this thesis, the person or group of persons participating in the game session and playing the game are referred to as players. The session is the occasion where the game is played but can contain other elements, such as a partial lecture, introduction to the game and debriefing. There can be multiple sessions if the game setup promotes it. The whole experience around the game and the session or multiple sessions are referred to as learning event(s).

1.2 Research process

The research process is twofold. The first phase is the model development through literature review. This literature review was carried out from an exploratory point of view. First, topics surrounding the benchmarking process were examined. Both academic literature and established business literature on benchmarking were examined. The goal was to establish the basis of what benchmarking is and how the process was to be carried out. Additionally, this verified the usability of the methodology for the evaluation.

After the basis for benchmarking was established, the different components which form a serious game were examined. From the literature on serious games, it became apparent

that the aspects from the session and the other material related to the game impact the suitability of the game. As there have been earlier serious game evaluation frameworks, these were also examined. Finally, project management was examined, as the literature on serious games highlighted the need for a solid connection to the learning goals from the game. To wrap up the first phase, the evaluated points were gathered into a single evaluation table, and the basis for evaluation scores was presented. The first phase covers steps one to three in figure 1.

Figure 1. The research process

The model developed in the first step is verified by benchmarking a set of serious games in the project management field. Following the setting of the benchmarking framework, the internet was searched through two search engines to find games to evaluate using the framework. The gathered list of games was drilled down based on availability in English. Also, purely tabletop games were cut from the list to clarify the evaluation. The remaining games were grouped based on the availability of a demo. During the game evaluations, the framework was evaluated alongside the games. After the evaluations were completed, the scores were totalled by evaluation area for each game. The benchmarking framework areas and points were also statistically evaluated between the games with demo and without demo options. The second phase covers steps four to six in figure 1.

The second chapter gathers the literature base for the model. The research methods and the data gathering process are covered in the third chapter. The findings from the evaluations of the games and the framework are presented in chapter four. These findings, alongside potential future points, are discussed further in chapter five. Finally, chapter six concludes the thesis.

2 LITERATURE REVIEW AND GATHERING OF THE FRAMEWORK

This chapter sets forward the development of the framework. First, the method of the literature review is presented. Following this, the literature on each subsection is examined. In each section, first, the literature surrounding each subsection is covered. In the final subsection of each section, the identified insights are gathered to form areas of the framework. To construct a baseline view of factors needed in the benchmarking framework for serious games, first, a broader look at benchmarking in general, benchmarking process and criteria is taken. Then the elements of teaching, games and serious games are examined to find the factors to evaluate. These gathered factors are compared to some existing serious games evaluation criteria to see which factors have proven beneficial earlier. As the serious games evaluated in this framework come from the project management sphere, some points on project management are finally considered. The final section gathers and formulates the developed framework together and forms together the benchmarking criteria used to evaluate the games. Here the evaluation points are also further detailed.

2.1 Literature review methods

At first, literature is examined to find the best benchmarking practices, how they are applied and which factors to consider. A literature review forms the foundation for the benchmarking process as a research methodology (Snyder 2019). After examining benchmarking practices, game components are examined to find the basis of comparison between the games. To examine the components of the serious games, a more narrative focus of the review is adopted. Using the narrative framework in the literature review provides a basis on which the framework gets constructed from the games components' perspective(Palmatier et al. 2018). As this is the focal goal of the literature review, the exploratory review was selected as a method of this literature review. In the exploratory literature review, the goal is to find what already exists in the domain of the topic at hand (Adams et al. 2007, p. 56).

Earlier serious games benchmarking models are examined against the benchmarking practices and the components of the serious game to see which areas might need more coverage. Based on the found gaps, the discovered best practices are applied to form a framework to evaluate the games. These gaps in each evaluation model are based on the results from the other benchmarking models as well as the exploratory literature review (Adams et al. 2007, p. 56).

2.2 Benchmarking

Benchmarking has had several different definitions depending on the point that is highlighted. One of the often-quoted definitions is that "*benchmarking is the search and implementation of industry best practices to achieve exceptional performance"* by Camp (1989, p.12). Anand and Kodali (2008) increased the detailing by describing benchmarking as "*a continuous analysis of strategies, functions, processes, products or services, performances, etc. compared within or between best-in-class organisations by obtaining information through appropriate data collection method, with the intention of assessing an organisation's current standards and thereby carry out self-improvement by implementing changes to scale or exceed those standards."* There has also been further refinement to this definition. Jetmarová (2011) highlighted why an organisation performs benchmarking as a methodology to competitor evaluation, market understanding, positioning, and opportunity identification. One of the critical factors when performing benchmarking is its goal-oriented nature, as a lack of goal is wasteful of resources used to perform the benchmarking(Zairi and Leonard 1994, p. 75). Broome and Quirk (2015) defined a different formulation of benchmarking. In their definition, they highlighted the core of benchmarking being the comparative assessment of one or many of the following aspects: Quality of conduct, how well the responsibilities are covered in said focal area; quality of design, the quality of underlying policies and rules; and the quality of outcomes, how well the goals are reached on specified areas.

The number of different benchmarking methodologies can be hard to gauge. Zairi and Leonard (1994, pp. 225–234) considered 14 different methodologies, while Anand and Kodali (2008) examined 35 different models to form their benchmarking framework for benchmarking models. The number of benchmarking models considered went up to 60

for Andersen and Moen (1999). Naturally, these listings overlapped. For example, the Xerox model was included in all of them. The Xerox model presented came from (Camp 1989, p. 17). It has been widely used as a baseline for benchmarking. For example, Anand and Kodali (2008) used it as the model to benchmark their model.

These numerous models come from different backgrounds. Baba, Mohd Yusof and Azhari (2006) notified of benchmarking models coming from two sources: academia and experts. The academic and research-based frameworks cover the academically developed benchmarking frameworks based on the researchers' research, knowledge, and experience. Academical models often had strengths when it came to coverage, but on the other hand, this increased their complexity and increased challenges in implementing and testing them. Alongside the frameworks presented in the paper, Anand and Kodali (2008) framework exemplify it with their 12 phases 54 steps universal model. Consultant and expert-based models base their models on the experiences and opinions of said consultants. While often more practical than the presented academic models, some also fall into the highly complicated category (Baba et al. 2006). Anand and Kodali (2008) added the group of organization-based models to cover models originating from specific organizations. They are often practical but so organization-specific that moving them to another environment can be challenging. Also, the variance is very high due to the business-specific nature. An example of an organization-specific model would be the Xerox model presented by Camp (1989, p. 13). Xerox model has often been used when evaluating other benchmarking models, such as Anand and Kodalis' (2008) universal benchmarking model. It was also highlighted by Zairi and Leonard (1994, p. 64) as one of the more comprehensible benchmarking models they evaluated, so for this thesis, this model is used as a basis when comparing benchmarking models.

2.2.1 Benchmarking model classifications

The definitions for benchmarking definitions presented earlier have been honed down to specify further different approaches to benchmarking. Fong and Cheng (1998) listed 13 different classifications for benchmarking methodologies based on their relation to each other, content and purpose of the relationship: Internal, Competitor, Industry, Generic, Global, Process, Functional, Performance, Strategic, Competitive and Collaborative.

However, these aspects have been widely used. The business world roots are shown in the differences in how these are named. For example, Zairi and Leonard (1994, p. 50) defined generic benchmarking as benchmarking of cross-functional processes, while Fong and Cheng (1998) defined it as Benchmarking beyond industrial boundaries. Anand and Kodali (2008) noted this same challenge in the benchmarking literature as Fong and Cheng (1998). Both suggested the same solution: stick with one and remain consistent.

Different approaches rose when examining the nature of the connection between the benchmarked parties. In internal benchmarking, the focus is on comparing the functions inside the organization against an internally defined benchmark, for example, another unit within the organization (Zairi and Leonard 1994, p. 49). Competitor benchmarking is a comparison of direct competitors to gain an advantage in competition or reduce the gap (Fong and Cheng 1998). Zairi and Leonards (1994, p. 34) definition is oriented more towards informing the benchmarker of the status of the competition. The industry-level benchmarking gets extended to include the non-competitors. Fong and Chengs' (1998) definition of generic benchmarking covers comparisons that extend beyond the comparing organizations within the industry, while global benchmarking covers benchmarking practices beyond country borders. As some of the lines between classifications in this sector have blurred as global operations have become more common, Anand and Kodali (2008) suggested simplifying the classification to consider only internal and external benchmarking, as all the others are subclasses of these two. Their suggestion also included the different purposes and varying benchmarking model content.

When examining the benchmarking content, the process benchmarking is first examined. Process benchmarking covers evaluating a specific work process or operating system (Fong and Cheng 1998). Functional benchmarking compares specified functions with best-in-industry or comparison organizations (Zairi and Leonard 1994, p. 48; Fong and Cheng 1998). Zairi and Leonard (1994, p. 50) extended this to include cross-functional processes with their form of generic benchmarking. Performance benchmarking covers the comparison of outcomes, reliability, results, etc. (Fong and Cheng 1998). Strategic benchmarking involves assessing strategy elements that impact it rather than product or process (Fong and Cheng 1998).

The relationship between the parties of benchmarking forms the third basis. (Fong and Cheng 1998) split this into two segments, collaborative and competitive. A collaborative benchmarking environment pushes toward forming a learning environment where the main goal is not to defeat the other partner but to enable development on both partners' sides. One of the methodologies for collaborative benchmarking in the private sector is the usage of Secure Multi-party Computation to hide the details of the participating organizations while enabling a push toward better practices industry-wide (Kerschbaum and Terzidis 2006). The benchmarker's goal is to gain superiority over the other benchmarking parties in a competitive relationship. In competitive benchmarking, the comparisons are made to the imitation of the other party, even if there is a partnership in place. (Fong and Cheng 1998)

2.2.2 Benchmarking process and data

There have been several different formulations of the benchmarking process. The number of steps and phases varies significantly between different models from different backgrounds, as Anand and Kodali (2008) noted. One of the most widely referenced models (for example, by (Zairi and Leonard 1994; Anand and Kodali 2008)) is the Xerox benchmarking model presented by Camp (1989, p.17). The Xerox model consists of four phases: planning, analysis, integration, and action, with ten steps in total, alongside the fifth area with two steps covering the maturity of the benchmarking practice. The number of phases and steps are not as integrally connected. For example, the AT&T model consists of the same number of total steps as Xerox, with 12 steps, but grouped into two phases (Zairi and Leonard, 1994, p. 61). Thou, when comparing the AT&T model to the Xerox model, the content of each phase could be described as a combination of the first two and lateral two phases, respectively. On the other end of the spectrum lies the universal benchmarking process Anand and Kodali, (2008) with 12 phases and 54 steps, the same number of phases as the Xerox model had steps in total. The analysed benchmarking models that formed the universal model proposed by Anand and Kodali had two to seven phases with five to 21 steps.

When examining the different models' steps, it can be hard to gauge steps one to one. For example, Zairi and Leonard (1994, p.57) As the first five steps set up the measuring and

data collection, these map to the first four steps on the xerox model as presented by Camp(1989, p.13). The same phenomenon can be seen when examining the list gathered by(Anand and Kodali 2008). Many of the identified steps either overlap or simply examine the different points of view on the topic. When examining the focal area of the benchmarking, the focus on process, product, or strategy did not strongly impact the number of steps (Zairi and Leonard, 1994, p.51-67).

In the Xerox model, the first step sets up the measured areas by establishing what needs to be benchmarked. Alongside the selection of focus areas comes the establishment of how these should be evaluated. When establishing these, Camp notes that it is essential to consider the point of view when selecting the evaluated metrics. The goal is to find the best practices, not simply measure the status quo. (Camp 1989, p. 16)

2.2.3 From the benchmarking literature toward the framework

The literature gathered covers the first step of the planning phase of the Xerox model as described by Camp (1989, p. 17). As the literature establishes evaluated areas to be benchmarked, benchmarked games are gathered and evaluated from steps two and three. The evaluated results are directed towards analysis and integration in certain aspects, as the gathered data gets analysed and communicated through the thesis. The benchmarking carried out in this thesis falls towards the competitive benchmarking where the benchmarked parties are external, as set by Fong and Cheng(1998). The benchmarking is also geared toward performance benchmarking in nature, as defined by Zairi and Leonard (1994).

When it comes to establishing the metrics to evaluate benchmarked games, selecting the focus of the benchmarking is important not to waste effort (Camp 1989; Zairi and Leonard 1994; Fong and Cheng 1998; Anand and Kodali 2008). In the following chapters, the aspects arising from serious games and projects management elements are further detailed. From the benchmarking literature, one of the performance and product benchmarking key areas is the costs associated with the product. As such, it gets its own key area in the benchmarking model. The cost evaluations consider the time required

from the staff for the session and the provided material, as it connects directly to the costs from the staff's time.

2.3 Which criteria to consider in benchmarking serious games

2.3.1 Game elements from serious games

As games consist of several different aspects, each of these areas is first examined to form an overview. The overall evaluation of video games has been dissected into several different components with different emphases on the aspects of games. The game elements are different aspects that form the basis for the game, such as the game's mechanics, UI and story. First, the actions are examined, then the environment in which the actions are taken, how these connect to form a story or experience and finally, how these different aspects impact the players' experience in the game.

The game mechanic is part of the game's rule system that single possible interaction with a possibility to occur in the game, be it specific interaction or a more generic event. A single game can have several mechanics, and the same mechanic can be across different games. (Lundgren and Björk 2003) *The genre* of the game can be defined based on the mechanics the game consists of and how those mechanics interact in the game (Björk et al. 2003). One of the core aspects of a game is a *gameplay loop*. Guardiola (2016) presented it as actions that the player takes in a game in verb format with minimal context. These actions are connected to one another in an example flow chart to enable an analysis of the actions the player takes during the gaming session and how these different actions connect.

Graphics is the level of visual detail included in the game. Aesthetics, on the other hand, cover the overall visual style. How pleasing a game is to the eye depends on the harmony of these factors with the context of said game (Hoenig 2005).

Alongside the game's visuals, audio can also be used to deliver context to the player's actions and provide guidance. The classification of different audio elements in video games comes from several aspects of how the sound is used in the game. Diegetic sounds

are coming from the game world, like the effect of an action. Non-diegetic sounds cover the sounds in-game that are not part of the game world, such as music in the soundtrack. (Jørgensen 2010) According to Jørgensen, Stockborger (2003) split these groups down into different sound objects, with the core of definition coming from the role the sound provides in the game.

Usability of a videogame has been defined as "*the degree to which a player is able to learn, control, and understand a game*" by Pinelle, Wong and Stach (2008). It has been further refined with the concept of *playability* to describe *"'a set of properties that describe the Player Experience using a specific game system whose main objective is to provide enjoyment and entertainment, by being credible and satisfying, when the player plays alone or in company', Also, Playability represents the degree to which specified users can achieve specified goals with effectiveness, efficiency and, especially, satisfaction and fun in a playable context of use"*(González Sánchez et al. 2011).

To evaluate playability, it has been broken down into the following attributes: Satisfaction, Learnability, Effectiveness, Immersion, Motivation, Emotion and Socialization (González Sánchez et al. 2011). An alternative point of view was presented by (Sweetser and Wyeth 2005) by examining video game elements to flow theory elements, which covers an examination of aspects that make experiences enjoyable. The elements were listed as concentration, the challenge to player skills, control, clear goals, feedback, immersion, and social interaction. The elements were further developed by examining the heuretics related to each aspect based on professional reviewers of the games alongside academic sources (Sweetser et al. 2012). Some of the elements in the game flow model map directly to aspects in the playability model by (González Sánchez et al. 2011), such as immersion and socialization, while others come closer in the evaluation criteria, such as control and challenges impacting the players' motivation. Playability is also essential for the achievement of learning goals. As noted by Cowley *et al.* (2013), a heavier mental workload weakened the learning results, but the specifics of the game have an impact on how the workload impacts the results. Zhonggen (2019) noted that the perceived ease of use of the game led to better results from the players, as it enabled them to focus on the content and enjoy playing the game.

2.3.2 Elements from serious games literature to consider in the benchmarking model

The literature on game aspects of serious games forms the basis for two of the evaluation areas: the mechanics and the game's flow. The mechanics of the game form the rules on which the game is built (Lundgren and Björk 2003). The rules are set to govern the actions the player can take in the game. These actions shape the gameplay loop, which is one of the ways of describing potential interaction chains inside a game(Guardiola 2016). As the goal of a serious game is to educate and entertain (M.Nazry and Romano 2017; Almeida and Simoes 2019), the mechanics and the gameplay loop have to support the core learning goals of the game. As the players learn the game and its rules, the game evolves to maintain a challenge and the interest it provides to the player.

As noted by González Sánchez, Gil Iranzo and Gutiérrez Vela (2011) and (Sweetser et al. 2012), the flow of the game and its playability provide one of the core areas in serious games when it comes to the evaluation of each of the games. The flow and playability create the player's draw to keep playing and focus in the session. As usability is one of the core aspects of a serious game due to its' possible limiting factor to the learning provided by the game(Cowley et al. 2013). As the playability and flow of the game could provide a full evaluation listing of a game, selected points are tied back towards the model, such as social aspects in the game or the session, the clarity of goals and UI, and the player's feedback receive. Alongside the feedback provided by the game, the content of the session in which the game is played provides its' own value for the learning experience(Zhonggen 2019). These aspects more strongly connected to the session rather than the game will be evaluated in their own evaluation area.

2.4 Earlier serious games benchmarking evaluation criteria

Mitgutsch and Alvarado (2012) proposed evaluation criteria set for serious games. The assessment model has an academic background with limited empirical implications, as they presented their model with two assessment cases. The evaluated aspects of the game in the purposeful, serious games assessment framework cover the following aspects of the game:

1. Purpose
2. Content & information
3. Game mechanics
4. Fiction & narrative
5. Aesthetics & graphics
6. Framing
7. Coherence and cohesiveness of the game systems.

Assessment of accessibility of video games for those with cognitive challenges, such as dyslexia or colour blindness, can also benefit the evaluation of video games, as the same factors that prevent others from interacting with the game raise impediments to others when it comes to learning. The mental workload in serious games also rises for those without these challenges, though it might not prevent them from participating in the game. Salvador-Ullauri, Acosta-Vargas and Luján-Mora (2020) broke down these impending aspects based on the Web Content Accessibility Guidelines. They listed the following considerations grouped as pairs:

- Interface
    - A limited number of objects
    - Enough size of objects
- User control
    - Enough time to interact with objects
    - Button to restart the game
- Identification with the game
    - Motivational elements
    - Configurable elements
- Feedback
    - Rewarding mechanisms
    - A simple text to show results
- Transmission of concepts
    - Representation of concepts early
    - Implicit educational content

In their discussion, Salvador-Ullauri, Acosta-Vargas and Luján-Mora (2020) noted that testers of the games pointed out that the evaluation criteria needed more input from motivational elements and rewards evaluation points. They also suggested removing the user control elements listed in their criterion, as all of the evaluated games cleared the criteria. While this academic model only had five testers, they covered ten games in their evaluation. It has also been noted that considerations of different learning styles can provide value to the whole learning process, as it can enhance the learning experience for the players (Zhonggen 2019).

Emmerich and Bockholt (2016) discussed their game design based on the constant evaluation of the anticipated effects of the game on the learners. The key principle contained the constant evaluation of the effects through experimentations during the game design loop. The design loop is the evaluation of the developed features among the design steps of the game. The impact was evaluated on set steps before starting the game design in the loop. As the evaluation of the impact is interwoven into the game's design loop, the game's anticipated impact is the core of the game's design. They noted that this connects to the purposeful, serious games assessment framework as suggested by Mitgutsch and Alvarado (2012). The application of this game design method has been limited, as Emmerich and Bockholt (2016) only noted three games where it had been thoroughly applied. Nevertheless, they noted in these three cases that the method had helped the game strengthen the impact and tone of the delivery methodology of the message.

2.4.1 Elements from serious games considered in the framework

When evaluating the elements from the game perspective, the evaluation areas incorporated into the benchmarking framework come from the game's mechanics alongside the game's flow and playability. Salvador-Ullauri, Acosta-Vargas and Luján-Mora (2020) noted when evaluating their model that player motivation and rewards are one of the missing elements of their model. This ties up with the playability model proposed by González Sánchez, Gil Iranzo and Gutiérrez Vela (2011). Clarity of controls is one of the areas where making the serious game more accessible supports also the learning goals. As noted by Cowley *et al.* (2013), the effort the player must overcome to participate in the game has the potential of weakening the learning results as effort gets

spent on struggling with the controls. The controls of the game are also one of the areas that Salvador-Ullauri, Acosta-Vargas and Luján-Mora (2020) listed in their benchmarking model.

From the mechanics' point of view, Emmerich and Bockholt (2016) noted that the learning goals of the game have to be interwoven into the way the game gets designed. The learning goals are expected to show a stronger connection between the mechanics of the game and the topic the game focuses on and the way the gameplay loop ties toward the learning goals. Thus, it gets its own position in the evaluation framework. Alongside the content in the game and its mechanics, the session and the instructions provided to the players have an impact as well, as noted by Zhonggen (2019).

2.5 Key subject areas of project management

A serious game without tying the game mechanics Emmerich and Bockholt (2016) or benchmarking without keeping the goal clearly in mind is a wasted effort (Camp 1989; Zairi and Leonard 1994; Anand and Kodali 2008). As the selected focus is the serious games related to project management evaluation criteria, the management areas and the project environment variables are covered in more detail.

Projects and project management has been defined in several different ways. Artto, Martinsuo and Kujala (2008, p.18) defined a project as *"a unique entity formed of complex and interrelated activities, having a predefined goal that must be completed by a specific time, within budget, and according to specification*". (Munns and Bjeirmi 1996) defined project management as "*the process of controlling the achievement of the project objectives".* Reaching the project objectives is to be achieved by applying tools and techniques that do not disturb the organisation's routine operations. At the same time, formatting a bit differently, at the core (Artto, Martinsuo and Kujala, 2008, p.26) definition lines up with this definition.

2.5.1 Core project variables

The core variables of projects have been defined in several different ways. The following two are gathered from project management certification bodies of knowledge (BoK) and study guides:

Table 1. core project management constraints or values based on two project management certification agencies grouped based on similarities

Certification	Project variables						Sources
PRINCE2	Time	Budget	Scope	Requirements	Quality	Risk	(Hinde 2012)
PMI PM BOK	Time	Budget	Scope	Requirements	Quality	Risk	(Project Management Institute 2017, sec. 2)

One popular presentation of project success is the iron triangle or triple constraint of project management. It is often presented as the project delivered inside the budget, timeframe, and quality requirements. These factors also form the basis for the listed project management variables (Hinde 2012; Project Management Institute 2017). The definition of a project from (Artto et al. 2008) included these factors without specifying them as the triple constraint. Alternative representations of the quality include scope, performance and requirements. However, they do not appear to be as strongly connected to the first two points as quality, based on analysis done on the connection of topics in abstracts of the text in the field. (Pollack et al. 2018) If the project is not able to deliver on one of these aspects, the project as a whole is deemed a failure (Bronte-stewart 2015).

There have been suggestions for adding different aspects to the Iron triangle consideration, such as sustainability (Ebbesen and Hope 2013), safety and dispute (Jha and Iyer 2007). Alongside, the soft pyramid has been used to describe the project teams' internal, contractors' and clients' satisfaction with the project to provide more subjective considerations for the iron triangles evaluation (Jha and Iyer 2007; Caccamese and Bagantini 2012). Sometimes soft pyramid is depicted as a pyramid constructed over the iron pyramid's base. These include considerations of the project's impact on the participants and society at larger. (Caccamese and Bagantini 2012)

There has also been a large amount of criticism towards the triple constraint and how viable it is as an evaluation method for projects. Along with traditional iron triangle evaluation, it is hard to find a successful megaproject, as the mega-scale scope and moving the goalposts make it hard to fit into the iron triangle model (Lehtonen 2014). Also, other considerations on the success of a project need to be considered in megaprojects. Such a disposition of nuclear waste requires safety to be of utmost importance, with lower considerations to costs and time (Lehtonen 2014). Alongside lack of consideration on the softer side can prove to be a challenge for the iron triangle (Caccamese and Bagantini 2012). Another key weakness of the iron triangle model is the lack of consideration of the project's business impact. For example, when the first Ford Taurus was developed, the project was marked as a failure as it was late. This latter became the most popular car in the USA. The second models project was successful based on iron triangles criteria, but the second model was a failure on the business side. (Shenhar 2004) Similar challenges impact ICT projects, with up to 80% reported failure rate on the iron triangle criteria (Bronte-stewart 2015). However, considerations of the eventual impact of the project are not considered in this evaluation. Alongside, the impact of one of the three points of the triangle is not linear. Doubling the budget is not able to halve the time spent, or minor changes in timeframe can have way more significant effects on the quality of the final result (Bronte-stewart 2015).

2.5.2 Alternative project success evaluation criteria

There have been several models presented for evaluation of project success models, such as the network mapping model (Lehtonen 2014), Strategic project leadership (Shenhar

2004) and project status model (Bronte-stewart 2015). Also, extending the iron triangle to include other aspects such as sustainability (Ebbesen and Hope 2013) or soft pyramid (Jha and Iyer 2007; Caccamese and Bagantini 2012) to cover the noticed shortcomings of the iron triangle model can be seen as evolving the model to better match the world in which projects operate.

The network mapping model starts with the assumption that for megaprojects, a better approach would be to start considering them as programs or projects. Here, the consideration would be more towards the openness of the constantly evolving system. (Lehtonen 2014) The evolution of systems pushes the megaprojects towards mode two strategies as defined by Regeer *et al.* (2009), where the challenge is inherently complex, requires knowledge and collaboration from a large variety of fields, and the impact is beyond the project's outputs, having an impact on the systems at a larger scale. To account for these factors, the network mapping model examines the accountability in the horizontal model across the participants in the project instead of vertically examining supervisor-subordinate relation (Lehtonen 2014).

The project status model builds on top of the iron triangle extending the considered points to cover the benefits, quality from stakeholders' perspectives and risks with the learning in the project. These are then mapped concerning the actual results, were over or under the estimates and by how much. These mappings can be evaluated throughout the project and after the completion to provide a more holistic view. (Bronte-stewart 2015)

2.5.3 Management areas of project management

Project management has been split into several different management areas to manage the presented project variables. When discussing these different areas, the certification institutes have different ways of breaking down the project management as a management subject area (Project Management Institute 2017) or by theme and phase as in PRINCE2 (Hinde 2012). In contrast, the mapping of the themes in PRINCE2 and the management practices in the Project management institutes' body of knowledge do not map one to one. For example, the business case theme of the PRINCE2(Hinde 2012, chap. 4) has no directly corresponding section in the PM BOK(Project Management Institute 2017).

Some of the base areas, such as Scope, can be found as a theme in PRINCE2 (Hinde 2012) and its own management area in the PM BoK(Project Management Institute 2017, sec. 5).

The PM BOK sets the management areas of project management as follows:

- Governing structure of a project(Project Management Institute 2017, sec. 4)
- Scope Management (Project Management Institute 2017, sec. 5)
- Schedule Management (Project Management Institute 2017, sec. 6)
- Cost management (Project Management Institute 2017, sec. 7)
- Quality management (Project Management Institute 2017, sec. 8)
- Resource management (Project Management Institute 2017, sec. 9)
- Communications management (Project Management Institute 2017, sec. 10)
- Risk management (Project Management Institute 2017, sec. 11)
- Procurement Management (Project Management Institute 2017, sec. 12)
- Stakeholder management(Project Management Institute 2017, sec. 13)

2.5.4 Elements from project management literature towards the evaluation framework

As the evaluated games focus on project management, examining some commonly considered success criteria for projects can provide ways of assessing the game. The strength of the connection between the evaluation and the feedback the player receives and the topic the session focuses around provides a connection to earlier discussed topics of mechanics and playability. Emmerich and Bockholt (2016) noted that the core learning goal has to be integrated into the gameplay from design onwards to provide a strong impact. When it comes to measuring success (Bronte-stewart 2015; Pollack et al. 2018) noted that a project is often considered a failure if it fails one of the three criteria of the iron triangle. There has been some criticism of the criteria and suggestions for better assessment frameworks (Shenhar 2004; Lehtonen 2014; Bronte-stewart 2015).

These project variables were also integrally included in the listings by the certification bodies of literature but with additional aspects that can be much harder to measure. Incorporating these games into the mechanics could provide additional value. Other

expansions have been suggested to the iron triangle (Jha and Iyer 2007; Caccamese and Bagantini 2012; Ebbesen and Hope 2013). Considerations are also given for the management areas included in the PM BoK.

2.6 The evaluation framework

The core aspects of the framework have been gathered in previous chapters from the literature surrounding benchmarking, serious games, game development and project management. The developed framework is presented in the following table 2. Each subject area consists of three evaluation items. Each of the evaluation items is shortly described in the box in italics. This is to support the evaluation. When referring to these evaluations in the latter breakdown, the evaluation areas apprevation and the number of evaluation item is used. The scale of zero to two points has been set based on the formulation of the evaluation criterion. The scoring breakdown is presented by each evaluation area in the following section. As the maximum score for the evaluation item is two, the maximum score for each evaluation area is six. The maximum total score is thirty points. The appendix 1 has shortened version of the evaluation framework for reference purposes.

Table 2. Evaluation areas and evaluation items for the serious games benchmarking model with short description.

Game session	GS 1. How long is the session/sessions *This evaluation targets the flexibility of the duration of a session or a round. Is the game possible to configure to fit the lesson plan, or would the course need to be formed around the game?*
	GS 2. How do social aspects tie to the session *Is there some interaction between the players, or does the game facilitate group work? Does the game actively hinder the possibility of group work.*
	GS 3. How the game ties into the learning event *Is there content in the game that can be used in the debriefing? Is there an initial round to get the players familiar with the game?*

Mechanics	Me 1. How gameplay loop connects to learning goals *As the gameplay loop describes the player's actions in the game, is there a connection between the actions and the learning goals?*
	Me 2. How the mechanics connect to the core gameplay loop *The mechanics set the rules in which the player takes action. How do these mechanics connect to the gameplay loop? Are there many mechanics with weaker connections, or are the core mechanics tightly connected to the core loop?*
	Me 3. How mechanics and loop and their impact evolve as the game progresses *As the session progresses, how do the mechanics evolve? Does the mechanics develop to more nuances, or are they turned obsolete by another mechanic in the game.*
Playability and flow of the game	PF 1. Clarity of controls and goals in-game *Clarity of controls is a key part of video games. If the controls and the goals are not clear, the player experiences frustration and can lose focus on the game's core content.*
	PF 2. How feedback provided during the session ties to learning goals and the clarity of feedback *Feedback is one of the core parts of serious games, as it provides learning opportunities. Is the game's feedback during the play in line with the learning goals? Does it provide pointers to where the player could succeed better? When it comes to the feedback and evaluations provided in the end, how well does it support the learning goals and how clearly the feedback is provided.*
	PF 3. Do the UI and aesthetic choices support learning *Does the UI get in the way of playing and draw focus from the game's learning goals? Alternatively, does the UI provide the information the player needs in a clear and concise form? Is all of the necessary information also available in the UI?*

Coverage of key aspects of project management	PM 1. Which of the management areas of project management are covered based on the listing in section 2.4.3 *The project management areas are listed in section 2.4.3. concise representation of project management subject areas. The goal of the evaluation is to see how many are considered in-game. Considerations can rise from the game's mechanics, setup or evaluation results.*
	PM 2. How do the selected management areas cover the stated learning goals of the game *As the stated learning goals are one of the key focal areas of the game, how many of the project management subject areas are included in the learning goals. These are evaluated separately from the mechanics to give a point of comparison.*
	PM 3. How many of the project variables were used in-game *Project variables are one of the ways to measure the projects' success. How many of these measurements are included in the game and in the feedback the player receives in-game. Both the evaluation of project results as well as the mechanics of the game connection to said variables are evaluated.*
Costs	Co 1. The time required from staff *The time required from the staff is one of the key resources which incur costs in teaching. It is also a limited resource for the course. As such, the time required from the staff should be evaluated between the options.*
	Co 2. The materials that are provided in the package *One of the areas that support the staff and reduce the time needed is the materials provided for the system. Suitable materials are also able to improve the results.*

	Co 3. How versatile are the project setup options *One of the costs incurring areas is fitting the game to the needs of the learning event. Higher flexibility can also ease fitting the game to different courses.*

2.6.1 Game Session

The first evaluated area concerns the way the game supports learning events and the structure of the gaming session. The area consists of three evaluation items: The flexibility of duration, the connection of social aspects to the session and the games ties to the learning event.

For the scoring on the first evaluation item, GS1, the goal was to evaluate the ease of fitting the game into lesson(s) plans from a time perspective. If the game had a set duration without the possibility to modify the duration, the scoring was zero points. This would include games with a set number of rounds or single session games which required a full day to complete. For the latter case, the zero came if it was impossible to pause or save the game state to continue later. For scoring one point, the game had a minimum of a two-hour session without the possibility of a shorter introductory round. If the game had multiple rounds, this point was evaluated based on a single round. To score two points in evaluation, the game had to have multiple different duration options as a possible setting parameter. An introductory round of the game as a tutorial was also considered a potential duration for a session, providing an option for a shorter round. This evaluation required that the number of rounds was also adjustable for multiple round games.

On the connection of social aspects to the session, GS2, zero points were awarded if the game actively inhibited group work. The examples of group work inhibiting design choices would include quick time events requiring instant reactions or set progress time without the possibility to pause. For one point, the game did not have the features preventing group work, but the game did not especially encourage group work either. For a score of two points, the game supported group work through multiplayer mode or while playing the game; there was plenty of time to discuss alternative action plans for each potential option.

As the game is part of a learning event, the connection of supporting material and content in the game is examined from the game's point of view, GS3. For a score of zero points, the game provided no feedback that could be discussed after the session. Also, the preparation material did not exist in the game or as additional material. For a score of one point, the game had feedback considering the results from the game, such as pass or fail and potentially a reason. The information required before the game is played is provided as a part of the game at the start of the session or as a separate file to be given to the students. For a total score of two points on this evaluation point, the game provided good debriefing content in its' report card, enabling discussion on how the score formed. Also, the information required for the game, such as the premise, was presented clearly in the game or as a separate file.

2.6.2 Mechanics

The game mechanics set up how the player interacts with the game inside the session. This is evaluated through three items: the gameplay loops connection to the learning goals, how the mechanics connect and shape the core loop, and how the mechanics and the loop evolve as the game progresses.

The first evaluation area is the connection of the gameplay loop to the selected point of view, Me1. For zero points, the project management learning goals are just a skin on top of the gameplay loop. To score one point, the gameplay loop supports the selected point of view and does not actively push away from the learning goals. To receive two points as the evaluation score, the game has to integrally be connected at the loop level to the learning goals; for example, part of the loop is managing backlog in an agile project.

When it comes to evaluating the connection between the mechanics and the gameplay loop, Me2, the game received zero points if there were no connection between the mechanics to the loop and learning goals. To score one point, the mechanics have an impact on the final score the players receive, but they cannot make different choices. For example, if the risks realize, they have no way later to remedy the situation potentially. To score two points, the players have ways to interact with the subject area, such as getting insurance against the risk or potentially choosing how the realization of risk impacts the

project results. Considerations are also given for the support of the loop or loops from secondary mechanics. For example, if keeping the budget is paramount in the simulation, how the schedule management supports it.

Thirdly the evolution of the mechanics and the loop, Me 3. For a score of zero, the mechanics could even recede during the game. To score one, the mechanics do not recede. There can be an introduction of new mechanics, or the previous mechanics could provide a new angle to the task. To score two in evolution, the mechanics evolve during the gameplay session(s). This can take the form of more complex mechanics as the game progresses. Alternatively, if previously existing mechanics gain new interactions with other mechanics or the new mechanics are introduced to provide new interactions with old mechanics.

2.6.3 Playability and flow of the game

Following the mechanics comes the flow of the game and the playability of the game. This is evaluated through three evaluation items: Clarity of controls in-game, the feedback the game provides and the aesthetic and UI choices of the game.

For the clarity of controls and goals evaluation, PF1, zero points would be if there was no explanation of the controls anywhere where the players could get this information. This includes that the controls are not explained in the session materials. Additionally, the goal of the game would not be defined. To score one point, core controls are explained in the tutorial or earlier in the session, but once the game starts, there is no way to access the information. The game's goal could be stated, but it could move so much that it becomes murky. For the score of two points, the information on controls is explained at the start, either through a tutorial or explained in the session. The information is also accessible during the gameplay through adaptive tooltips or players' handbook in or outside the game. The goals are clearly stated in the game. If the goals change, this is also communicated clearly with pointers on how they changed. The tutorial can also be a warm-up game or round completed before the session to get the players familiar with the topic.

When it comes to evaluating the feedback the game provides, PF2 can be split into two evaluations. For these evaluations, the game must reach the said score on both evaluation angles. First, the score given to the player at the end of the session, how it is formulated and is it possible to use it as an example for grading. The second half comes from the feedback during the game. For a zero point score, there is no feedback during the session. For example, if a risk in the game is realised, it has no impact on the visualized costs, timetable progress, etc. For one point score, the feedback comes inside the game, for example, as a table of values. For a two-point score, the feedback is shown during the playing, and if the event impacts later decisions, it is noted to the player.

As for the clarity and support provided by the UI, PF3, the score can be gathered from the features the UI has. Is it possible to modify the UI to increase the font size or zoom in on text. Are the icons clear enough for legibility. Does the game have colour blind options if the colouration has the potential to cause issues for a person without full-colour sight. Furthermore, finally how is the information the player needs available and formatted in the UI. Here, the scale is formed based on the evaluation points, half for each point. This result then gets rounded up.

2.6.4 Coverage of key aspects of project management

The evaluation focuses on serious games in the project management realm. The learning goals are aligned with project management base aspects as collected from certification bodies of knowledge and academic literature. This is evaluated through three evaluation items: Which of the project management areas from Project Management Institutes BoK are covered (2017), how these project management areas are covered in the stated learning goals and which of the project management variables in either PRINCE2 (2012) or PMI BoK (2017)

The evaluation of key project management areas, PM1, is tied in parts to the mechanics' evaluation, as tying the core learning goals to the evaluated topic improves the learning goals. When it comes to evaluating the coverage of key project management areas, for a zero score, the game would only contain one to four, and none of them would be part of the game's core mechanics. For one point score, three to five management areas get

covered. At least two of the management areas form part of the core mechanics. For the highest score of two, more than four management areas must be present in the game. Out of these, at least three must be part of the core mechanics, and other management areas must have a meaningful supporting role in the game.

The second evaluation item is how these variables are connected to the game, PM2. To score a zero, the project variables are merely added as a finishing touch; in the example at the final scoring, the risk levels approached are considered, but the risks never impacted the play during the gameplay. For a score of one point, more core project variables are considered inside the game in some game elements, such as pop-ups, but the core gameplay loop ignores their impact entirely. This could take the form of risks without impacting the project time, quality or budget; it just creates a pop-up. To get two as the score, the game's main gameplay loop has to roll around these management areas.

The third evaluation item is the usage of project variables in-game and in evaluation, PM3. To get a zero score, these are not included at all. Alternatively, if one to two is included, they do not have a noticeable part in the gameplay or evaluation. To score one, the variables are included to some extent. They might not be part of the game's core, but they are there. To score two, the variables must directly impact the game as it is played. They must also be used in final feedback, such as a scorecard or an explanation if the result is failed.

2.6.5 Costs

Finally, the costs of the game and the session presents. These are evaluated through three evaluation items: the time investment required from the staff, the materials provided with the game and the versatility of the project setup options.

When it comes to the scoring on the time required from staff, Co1, to receive a score of zero points, the game would require extensive configuring and familiarity with the game to set up and run the session. Also, the game could require many staff members in session to guide and aid the game. For a score of one point, the configuring can take time, but it is not extensive. Also, some instructional material exists. For a total score of two points,

the game comes configured. Also, there are instructions for the game, and the game does not require a large number of staff to run.

The second costs related evaluation point is the material provided with the game, Co2. Here scoring of zero points means no materials. To reach a score of one point, the game must have at minimum instructions or scenario briefings for students or debriefing material. For a score of two points, briefings materials are required with instructions or lesson plans for running the game. Also, debriefing material and introductory material for students are required.

For the third evaluation, Co3, the versatility of the game is evaluated. If the game is possible to use across different teaching scenarios in the university, the adaptation costs of the game can come lower. Here the zero point score is one narrowly focused scenario without possibilities for customization. One point is given if the game has enough broad focus to enable it to be used to teach multiple concepts. Alternatively, it can also have other scenarios for this. To reach a score of two points, the game has to have either customization options, multiple varied scenarios with different focus areas, or the publisher has a linked game focusing on other aspects.

3 RESEARCH METHODS AND THE PROCESS

This section presents the research process and the used research methods of this thesis. First, the construction of the model is examined based on the literature review. Following the literature review, the empirical methods are covered, and finally, the data collection is discussed.

3.1 Research design

Following the building of the framework, the games to be benchmarked are gathered into a list for evaluation. Selection criteria for the games are focused on project management and availability in English. Once the gathered games are evaluated against the framework, the scores are compared between each evaluation point, category, and total. These are analysed as groupings of games based on the availability of assessment material.

By adapting interpretative, comparative concept analysis to the benchmarking practice, the benchmarking methods are used as a basis of action research (Kyrö 2004). Action research is a cyclical research process, which has four phases "*starting with the recognition of the problem, then planning the action, proceeding to carrying this out and finally evaluating the results obtained*" (Kyrö 2004). Atweh, Kemmis and Weeks (1998, p. 21,22) defined the phases as planning, acting and observing, reflecting and re-planning, adding the notation that the process is cyclical in nature. This thesis falls into the education-focused branches of action research practices by focusing on serious games.

Benchmarking has been selected as a methodology as it can be used to find best practices (Camp 1989, p. xi). While benchmarking has its roots in the corporate world, it can be adopted into academic research methodology. As explained by Zairi and Leonard (1994, p. 86), the cyclical nature of benchmarking through planning and setting standards, performing and measuring and conducting improvements can be seen as similar process as described by Atweh Kemmis and Weeks (1998, p. 21,22) described above. Using theoretical work to establish the evaluation framework pushes the benchmarking practice towards the action research subcategory (Kyrö 2004).

Alongside evaluating the games, the framework gets evaluated on each of the evaluation items to examine their role in the framework. After the evaluations, the framework is assessed based on the scoring given to each of the points. All of the games were evaluated by the author. For two of the games, two additional evaluators were used to see how the framework seemed to a person who had no prior experience with project management but was familiar with simulations and games. For good evaluations of the usability of a benchmarking model, there should be five evaluators, but using three evaluators reaches around 67% of the insights available in the model (Nielsen 2000). While this is a weakness of the model evaluations, broader testing could provide additional insights on the tuning and balancing of the model.

3.2 Data collection

The evaluated games or simulations were gathered through Google, and DuckDuckGo searches between 27.12.-21 and 30.12.-21, rounding out round 10.1-14.1.2022. The following phrases were used in the search. They were selected based on the keywords rising from the focus area of project management combined with terms related to serious games and simulations.

1. project management simulation game
2. project simulation
3. project management serious game
4. project management educational game

The search was carried out also in the fifth and sixth search terms. However, these provided poor results, as a project is a common word in video games for entertainment, and those titles dominated the results with related conversations on the internet. Project management games also provided many articles and blog posts on project management in game projects, alongside blogposts on how serious games can be used in project management training.

5. project game
6. project management game

The first five pages of the results were considered. A total of 21 games were discovered after removing the duplicates from the search results. As the search phrases were in English, most appearing games were also available in English. Two games were removed from the list as they were not available in English. After the language filtering, 19 games formed the initial sample.

Out of the 19 games, four were tabletop games and were left out as evaluation between computer and tabletop games would have been challenging. The evaluation would have been challenging due to the material available on the games.

Out of the remaining 15 games, five had freely available demos online, with an additional two having contact forms for requesting a demo. These were sent out on 22.02.2022. Out of the games that did not have demos available for their game, one had materials online. Two additional games had materials available through a contact form sent out on 24.2.2022. One of the games with form filling to get access materials, form never seemingly processed. The resending was attempted again on 3.3.2022 without success.

If the publisher had several games, they were evaluated as scenario options, as they often shared marketing material and were similar in approaches. Also, the differences between games from the same publisher were smaller than those between different scenarios for some games. This resulted in two publishers having their games evaluations combined. This elimination combined three games into a single evaluation for one of the publishers and two to one for the other case. In the end, five games were evaluated based on their demo, marketing material and instructional material available and seven games were evaluated based on their instructional or marketing material.

3.3 Evaluated games

Here the evaluated games are described in more detail. The grouping is based on what material was available for each game. As the data between these groups is somewhat different, they are analysed separately. The games are presented in alphabetical order in the following sections.

The evaluated games are described in the following table 3. The order in which the games are presented is the same order in which the games were found.

Table 3. The evaluated games

Name	Number of scenarios	Available material	Reference
Simultrain	7	Demo, marketing website, presentation video, instructional material	(STS Sauter Training & Simulation SA 2022)
Cesim	1	Marketing website, presentation video	(Cesim Oy 2022)
Simproject	2	Marketing website, instructor material, presentation video	(Simulation Powered Learning 2021)
Online complex project simulation	1	Marketing website, video presentation	(International Centre for Complex Project Management 2022)
ThatPMGame	1	Free online game, the material on the website	(GamesByRobc 2022)
AbleSim	3	Demo, marketing website, video presentation	(AbleSim 2022)

Project Management Game	3	Marketing material, screenshots, instructional material	(AlbaSim 2020)
Countdown	1	Marketing website	(Paradigm Learning 2022)
Business simulations -project management simulation	3	Marketing website, presenting video,	(Business Simulations Ltd 2022)
MIT Project Management Game	3	Demo, Marketing website, White paper	(MIT Sloan School of Management 2022)
Doublemasters	1	Marketing website, instructor material,	(Double Masters 2022)
The crowd training	1	Demo, marketing material	(The Crowd Training 2018)

3.3.1 Games with demo possibility or free version

AbleSim is a project management simulation game that focuses on managing the task net through scheduling resources for each of the tasks. The game has two language options with up to three scenario alternatives depending on the language. As the core is built on assigning resources to tasks, the gameplay focuses on giving human resources for each

task day by day on a weekly basis. For the demo, the player plays through a network where one task is performed concurrently with two tasks that come one after another.

MIT project management game is a project simulation with a two-phased project model in the three provided scenarios. The provided scenarios cover construction, product and software projects. There is also a custom option for setting up custom scenarios provided for educational partners. The player is the project manager in the game, guiding it through the phases with a reward if the player delivers it early and under budget. The management comes in from adjusting the staffing, quality requirements, pressure with time and quality and scope management by cutting or approving features from the final deliverable. Once the game is completed, the player is presented with a final evaluation of the results. Additional materials are available for educational purposes. The game has also partnered with a training simulators provider to provide commercial training and materials options.

Simultrains project management simulation puts the player in the project management position in their office. There are seven different scenarios available, ranging from event planning to agile project management. The simulations have options for customizing the duration from play speed to risks and scope changes. In the demo, the player can plan the project and play through the project's first week. Alongside the demo, their website contains trainers and students' materials with recommendations on how the simulation is run.

ThatPMGame is an online project game where the player assigns resources to a randomly generated five-phase task net. The assignable resources also have a randomly generated profile with three options for both speed and cost. In the middle of the simulation, the player has an option to accept a random event that can be positive or negative. During this pause, it is also possible to change the manning of the operation. The instructions for the game are included on the site, along with a side discussion forum for the game. A single round takes about five minutes.

TheCrowdTraining has a different approach to the serious game, as it would be more accurate to describe it as a virtual lesson instead of a simulation. The game contains training material for certifications and the project management profession. Different

game sections contain quizzes to test knowledge for certification exams, information on the application and similar topics. These do not contain a direct way to interact with other students.

3.3.2 Games evaluated with sales material

Business simulations have three games targeting the project management area: Crew, Crisis and Edge. Alongside targeting project management, the games have a side of high performing teams management, leadership and change management. From the project management side, the games focus on managing the effective use of the team, decision making, change management and time management. Each game is provided with its own video and list of functionalities directing the games to different types of events.

Countdown is a project management game by Paradigm learning. The game's goal is to provide the players the core principles of project management through teaching concepts, tools, techniques, and behaviours of effective project management—the project is based on the agile model. The game is designed as a classroom-based experience.

Double masters simulations provide project management simulations for their students. They provide users guides with descriptions of how the sessions are run on their sites. Additionally, comparisons are provided between the versions of the game. In the game, the player manages a project from start to end, making decisions, scheduling and budgeting, controlling the project, and understanding human resources and their impact on the project. There is also a short video describing the game.

Online Complex Project Simulation is a project management simulation by ICCPM. In their demo video, they present new stadium project management. The simulation describes the key topics as advanced project leadership, risk management, strategy implementation, and stakeholder management. The game is sold on the site as part of a workshop.

Cesims Project is a project management simulation with the addition of program management. The players work together as project managers of individual projects to bring the whole program to a goal. The players can decide on prioritising tasks, usage of

overtime, preparation for changes, outsourcing and staffing of the projects. The goal for individual projects is to complete them within budget with as low costs as possible, on time and with as good quality as possible.

Project Management Game is a simulation by AlbaSim. The player is a project manager for a virtual project in the game. The simulation covers the whole project, from selling it to clients and reporting to the steering committee to solution searching and documentation. The scenarios run from new product development and public sector project management to non-profit awareness camping management. The site provides screenshots of the game and manuals, and basic information on the game.

SimProject is a project management simulation by Simulation powered learning. They also have SimAgile named game that focuses on managing an agile project. For the SimProject game, the players manage an 11-week project. They have ten potential team members to choose from for their team. The game has been used as a classroom activity and standalone home assignment. For the agile versions, the project managers manage sprint planning, backlog refining, changes to the project, daily standups, etc. They also have a presentation video for the SimAgile.

3.4 Data analysis

The data analysis is covered in two parts. First, the analysis for the evaluations of the games is described. In the second subsection, the analysis of the framework itself is described.

3.4.1 Evaluations for the games

When the games were evaluated, each game was given a score based on the evaluation criteria outlined in chapter 2.6 and its subsections. When assessing the games, the marketing material was accepted at face value, as long as the material was related to the game at hand.

The evaluation items were given a descriptive statistical analysis. When evaluating the framework, the evaluation points were calculated with a mean value and variance based

on each of the evaluation items. These scores were also compared across different evaluation groupings, for example, the type of material available. To form a combined score for evaluation purposes of the games, the scores were added together. The evaluated scores were also totalled based on categories to provide comparison points between games.

3.4.2 Evaluations for the benchmarking framework

When evaluating the games, a note was also taken on each question on the difficulty level of applying said benchmarking criteria for the potential fine-tuning of the evaluation framework. The scale ran from zero points for very difficult to evaluate to two points for easy to evaluate.

The evaluation items were also evaluated on insightfulness. The framework was evaluated based on a zero to two-point scale, where zero meant the question at hand was not applicable for the game's evaluation, one if the question applied to this point and two if the fit was excellent or provided interesting insights.

The same descriptive statistical evaluation was carried out on the evaluations for the evaluation framework as was carried out against the evaluations of the games. The means and variances were compared across the game groupings at the evaluation area and item levels.

After the evaluation of the game was carried out with the evaluation of the benchmarking framework, final consideration was given to the points of the benchmarking question; for example, if several of the framework points focused on the same aspects, which one targeted the evaluation goals better.

4 FINDINGS AND ASSESSMENT OF THE RESULTS

In this chapter, the results from the evaluations are presented. In the descriptive statistical evaluations, the data is split into two groups based on the availability of the material. If the game had an online demo available for evaluation or accessible online game mode, the game is in group 1: games with demo possibility. Rest of the games formed group 2: games without demo. The evaluation scoring of each game by evaluation item is presented in appendix 2. First, the data is evaluated across the evaluation areas of the games, which is further drilled down to evaluation items.

In the following section, 4.1, the evaluations of the games are examined. In the first subsection, the scoring of the games is examined statically by evaluation area and item. Following the statistical evaluation, the next subsection presents the scoring for each game by the total the game received and the total scores for the evaluation areas.

After the evaluations of the games, the next section, 4.2, presents the evaluation results for the framework. The first subsection presents the ease of evaluation results statistics with a breakdown by evaluation area and item. These statistical evaluations are compared between the evaluation groups. The second subsection focuses on the insightfulness of each evaluation area and item. The statistics are presented as well between the evaluation groups.

4.1 Evaluations of the games

To answer research question two, the evaluation results of individual games are presented. First, the statistics of these evaluation areas are presented to examine the fit of the final sample against these evaluations. The distribution of scores received is presented for the research question one's analysis. Finally, the game scores are presented to assess research question one.

4.1.1 Statistics on the scorings

The evaluated games had the highest mean evaluation by evaluation item in the project management evaluation area. It also had the lowest variance. The mechanics' evaluation area had the lowest mean score for the evaluation items. The variance was the highest for the costs of the session evaluation items. Figure 2 presents the means and variances of evaluation items by evaluation area. The same colour scheme is used for the groups for the rest of the thesis.

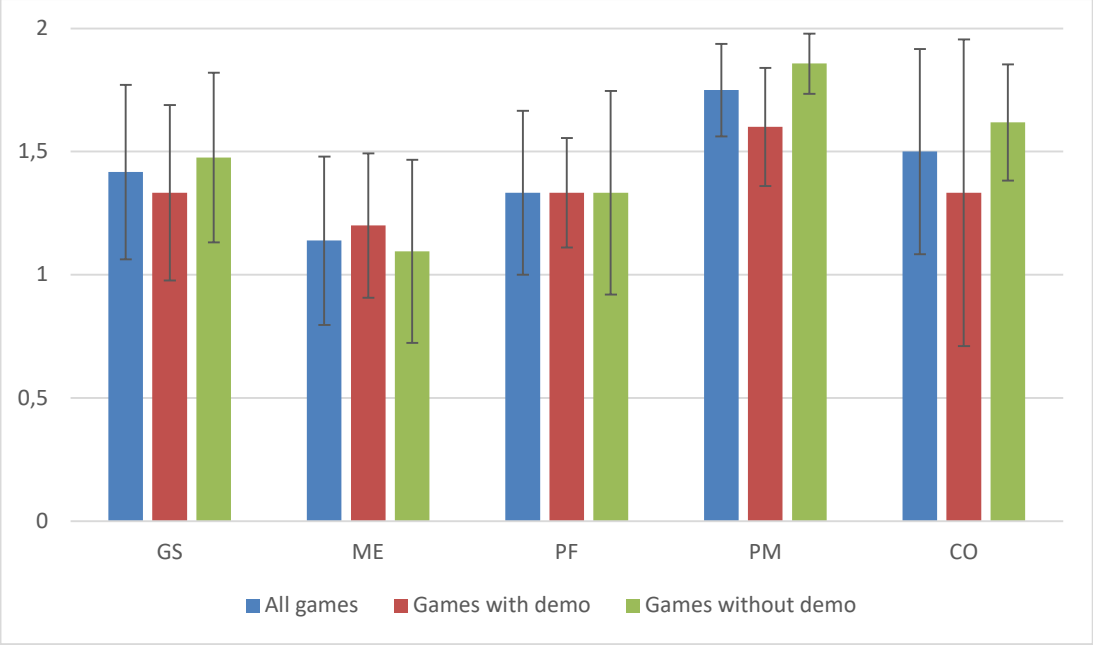

Figure 2. Means and variances of each evaluation area from the whole set and broken down on the demo availability.

For the game session evaluation area, GS, the demo possibility games had a lower mean and slightly larger variance than the games without a demo option. When evaluating the mechanics of the game, games with demo options had a mean score higher than average. Interestingly, the sets had the same mean score from the playability area, but the variances varied more than for the previous evaluation areas. The project management subject area had the highest average across all of the evaluation areas. Here the games without demo possibility had a higher average than those with demos available. Finally, on the costs, games with demos scored lower on average than games without a demo but had nearly

three-time higher variance than the games without demo possibility. These are presented in figure 2.

Figure 3 presents the means and variances from each evaluation point for all games. The games' lowest evaluated score was the evolution of the mechanics and the gameplay loop during the session, Me3, with the way the mechanics connect and shape the core loop, Me2 alongside the flexibility of the session(s) duration, GS1. While the mechanics had the lowest mean evaluation score, the lowest variance came from the flexibility of the session evaluation scores. The variance of the mechanics' connection was double to the variance of the session flexibility. The highest mean came from the connection of the session and the game, GS3, and the connection of the stated project management subject areas to the project management areas found in the demo or other marketing material and instructions, PM2. This also had one of the lowest variances with the flexibility of the duration. Interestingly the highest variance was on the question of how the social aspects tie into the game, GS2.

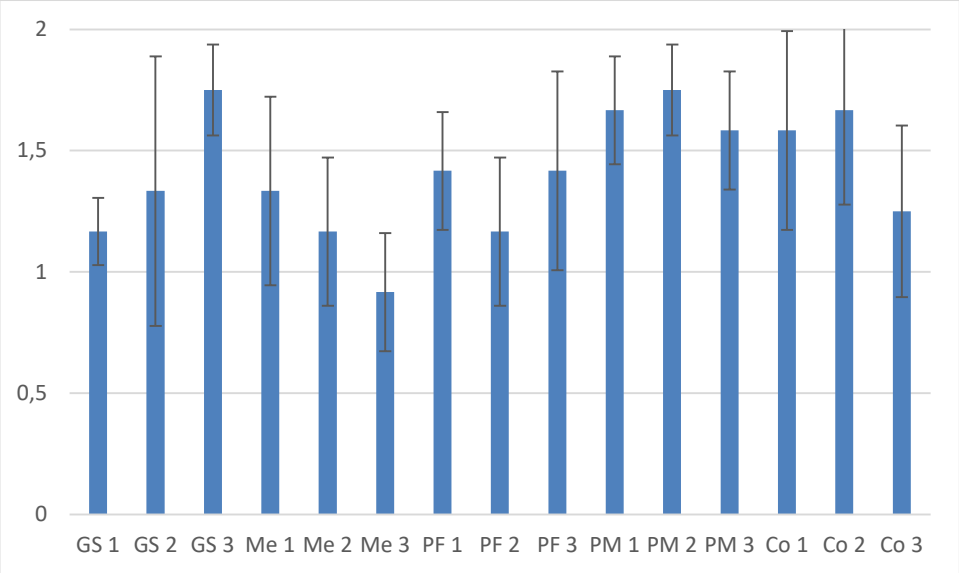

Figure 3. The means and variances of the evaluated games

When comparing the scores of the games with demo possibilities, the scores aligned with the means of the sets. The higher mean of the mechanics' evaluation than the games

without demo possibility came mainly from scores on the connection of gameplay loop and the evolution of mechanics. The largest difference in the means is in the coverage of the project management areas with the stated learning goals, PM2. When it comes to variances with these two data sets, typically, the demo games have more significant variances than the variance with all of the evaluated games. The mechanics, the gameplay loop evaluations, feedback and UI are the only areas where this pattern is broken, as the variance is larger with the whole data set. The means and variances for the games with demos are presented in figure 4.

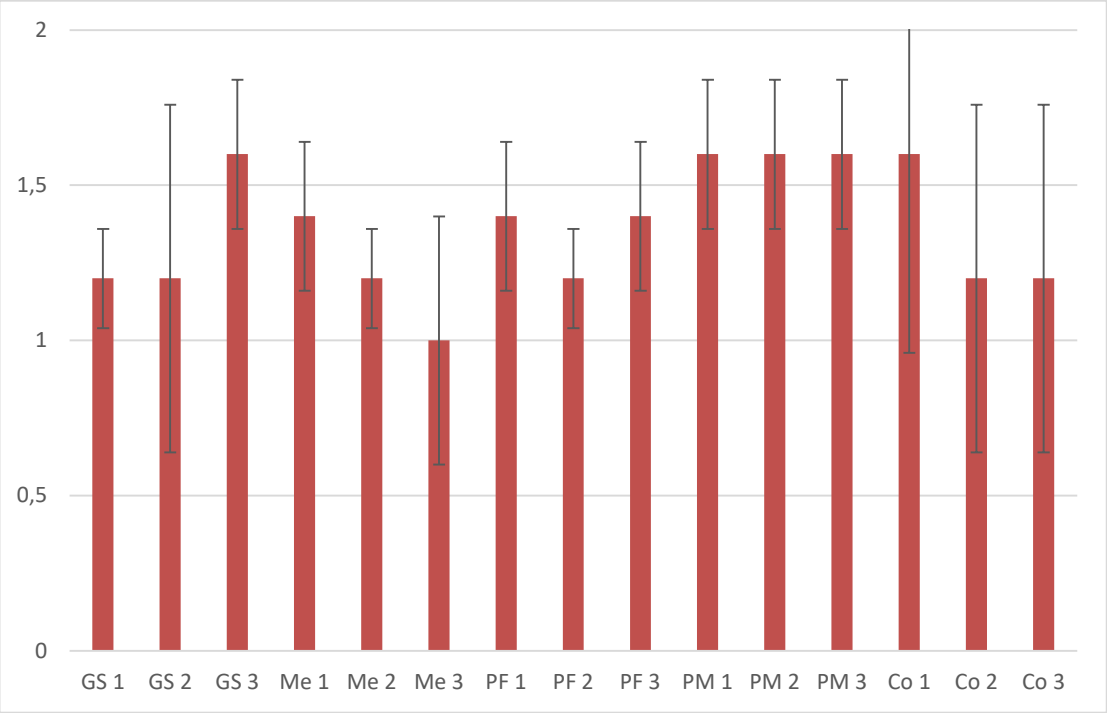

Figure 4. The means and variances of the games with demos

The highest scores came from the project management connected area for the games without a demo or free testing mode. They also had one of the lowest variances as well. These questions were concerning the coverage of the stated learning goals, PM2, and the quality of the provided additional material, such as case descriptions and debriefing material, Co2. For several of the games, there was also a possibility of additional services for several of the games, such as the company hosting the game as a workshop in the example. These options were often more targeted toward business customers rather than

academically focused ones. Low variances were also found in the evaluations of the connection to the learning event, the evolution of the gameplay loop and mechanics as the game progressed and the coverage of the project management areas based on the PM BoK listing. Several of the games described using this said listing as a guideline in their game design as well. The evaluation points which had the lowest variability also had the highest mean scores, except for the evolution of the mechanics, where the mean value was the lowest. The means and variances are presented in figure 5.

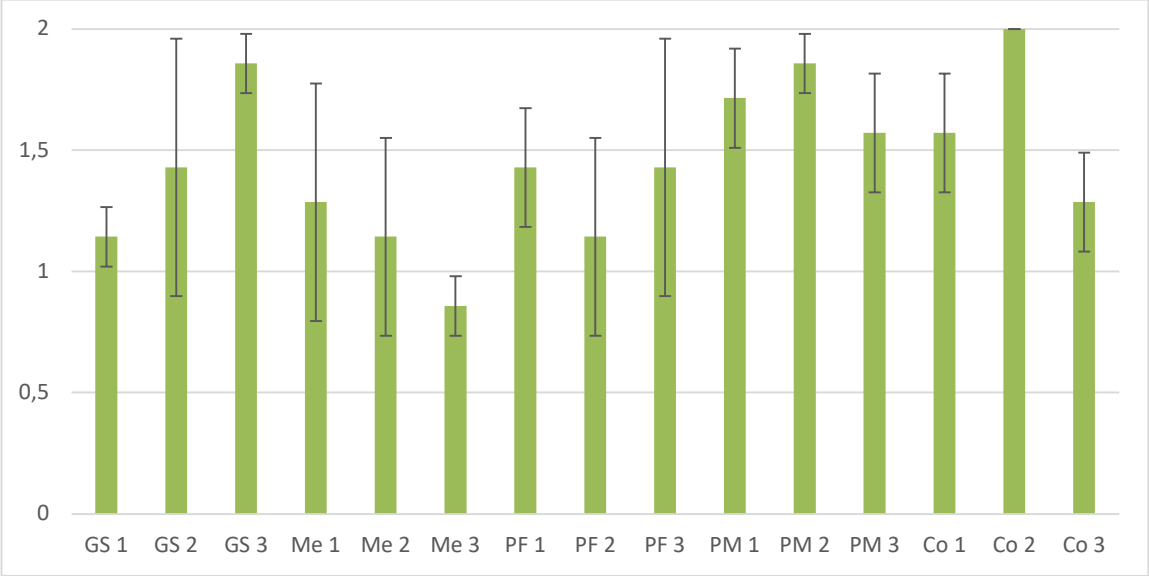

Figure 5. The means and variances of the games without a demo option

4.1.2 Total scores of each area by game

Figure 6. presents the number of times each possible evaluation score appeared in each category. Across the dataset, scores of two and one were given 85 and 84 times, respectively. To receive an evaluation of zero was the least common; it came up only 11 times.

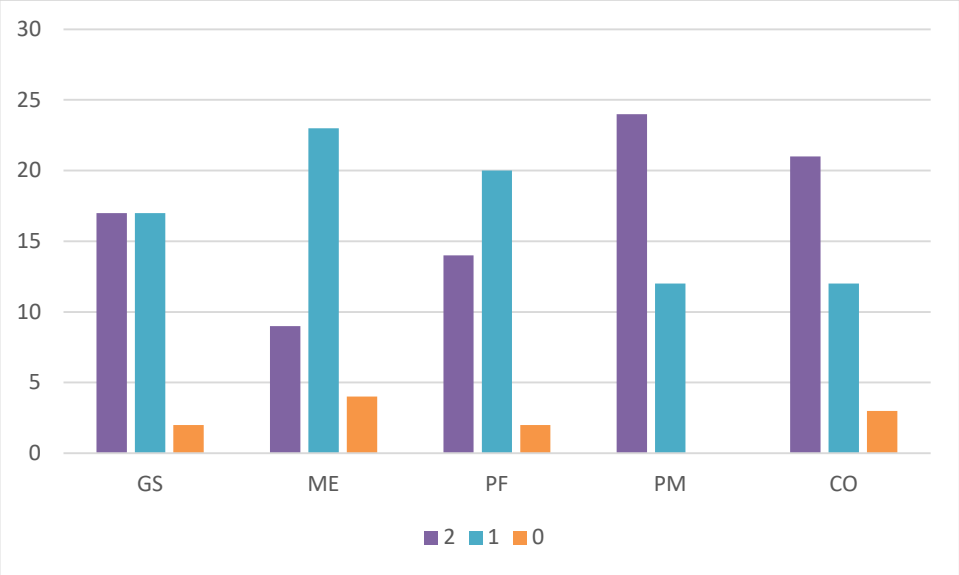

Figure 6. The number of times a score was given within each category

In the game session area, GS, scores one and two were given 17 times. For the mechanics' evaluation area, Me, the most common score was one. It was three times more likely for a game to receive score one or zero than two in this evaluation area. While one was the most common score also for the playability and flow of the game area, PF, the difference was less stark. In the project management area, the most common rating was two, being twice more likely than a score of one. There were also no zeroes in the project management area. Two was also the most common score for the costs area. Here some of the games received the evaluation of zero.

In table 4, the total evaluation score for each subject area is presented alongside the total score across all subject areas. These listings are presented in order of descending total score. The total scores range was from 26 to 12 when the maximum total is 30. For each individual area, the maximum total is six. This is reached by having the evaluation of two in each of the evaluations points. The full data is available in appendix 2

Table 4. Total score of each evaluation area and the total overall score

Evaluated game number	GS	Me	PF	PM	Co	Total
1	4	6	5	6	5	26
2	5	5	5	6	4	25
7	5	4	6	5	5	25
3	6	3	4	5	6	24
4	5	4	4	6	5	24
9	3	4	5	5	5	22
6	5	3	4	4	5	21
10	5	4	4	5	2	20
11	3	3	3	6	5	20
12	4	2	3	6	5	20
5	2	3	4	3	3	15
8	4	0	1	3	4	12

Across the evaluated games, the lowest score a game received was 12. To reach the higher half of the evaluation results, the game had to score a minimum of 22 points across all categories. The top score was reached in all of the categories. Most commonly, the score

of six was reached in the project management connected evaluations, a total of five times. The other areas each received an evaluation of six once each. The most common evaluation total for a subject area was five, with a total of twenty hits. It was also the most common score in the costs and game session categories. A score of four and three was approximately as common, having 15 and 11 hits, respectively. They were the most common score in the mechanics' category. Scoring of two and one were the rarest; only playability had a game to evaluate to 1, while game session, mechanics and costs had a game evaluate to 2.

4.2 Evaluations of the framework

Alongside evaluating the games, the framework itself was under evaluation. This was examined empirically to address the first research question on the suitability of the evaluation areas for the evaluation of serious games. Here the gathered data covers how difficult the formation of the score was during the evaluation and did the evaluation point seem to be relevant or brought out new information on the game.

4.2.1 Ease of evaluation

Across all evaluation areas, games with demo options were rated easier to evaluate than those without the demo. The difference between the evaluated ease of evaluation was at its' largest when comparing the project management related area. The closest the ease of evaluation score was between the groups on the game session related area. The mechanics were the second-largest difference between the groups. When it comes to the variance of the dataset, the games with demo options had higher variances than the games without demos on the mechanics and playability and flow-related areas. The most significant difference between the variances was found in the game session area, GS, where the games without demos had around double the variance of the games with demos. The means and variances for the ease of evaluation are presented in figure 7.

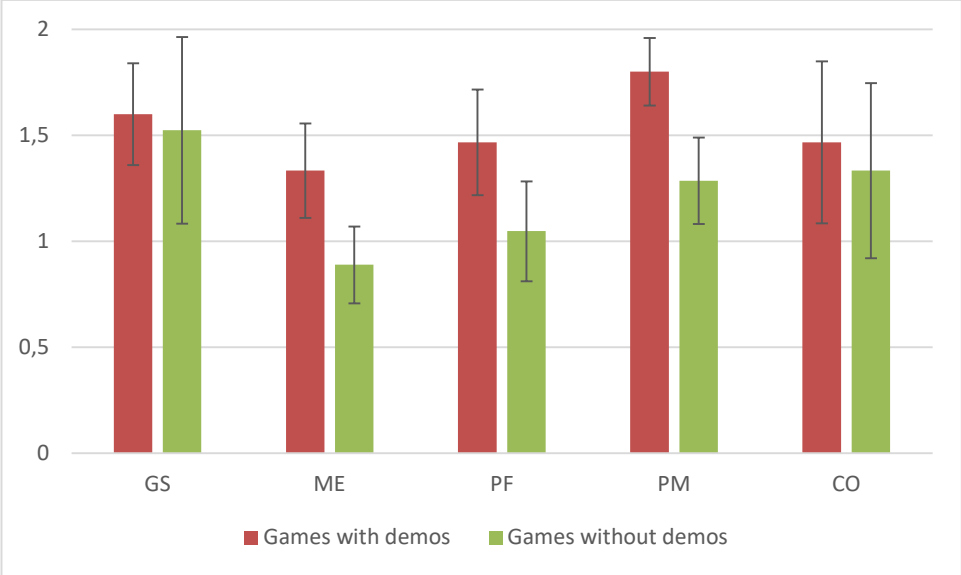

Figure 7. The means and variances of ease of evaluation across frameworks areas.

The same trend continues when it comes to the evaluation points inside the subject area evaluations. Only in the game's connection to the learning event and the materials provided evaluation items were the games without demos were rated easier to evaluate than the games with demos. The UI-related assessments were the only item where a single evaluation item had a much larger difference between the ease of evaluation and the mean ease of evaluation of the area. The variances between the evaluations on the individual points had a large difference to the grouped variance of each area, and between the groups. This is presented in figure 8.

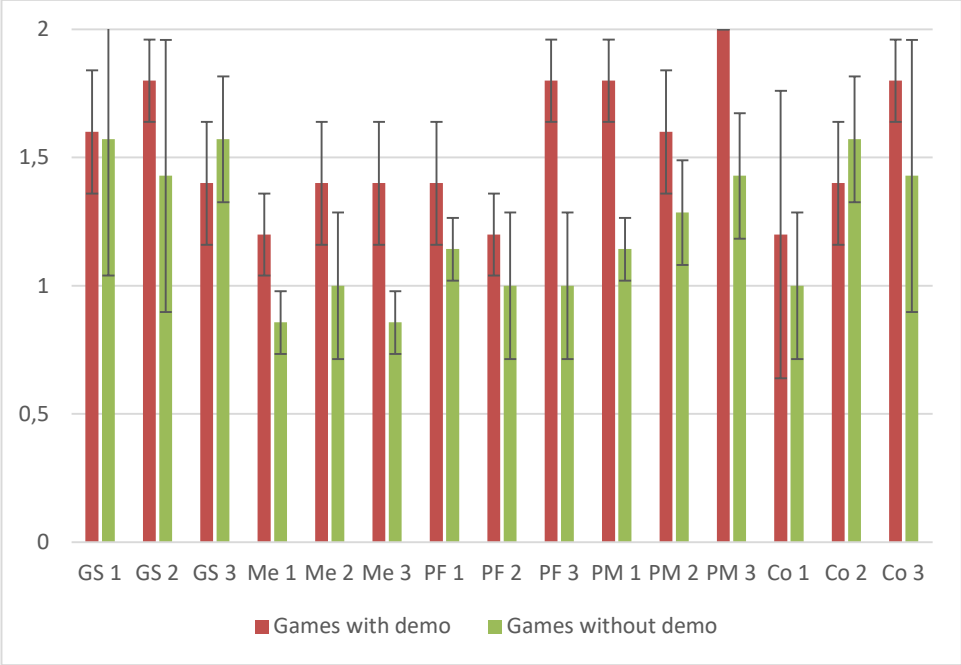

Figure 8. Means and variances of ease of evaluations for evaluation point broken down on the availability of demo.

4.2.2 The insightfulness of evaluation item

When it comes to evaluating the insightfulness of each evaluation point, the most considerable difference was in the evaluation of the mechanics' area. Here the games with demos had over half a point higher mean than those without demos. Across the other evaluation areas, the largest difference of means was 0.1. Game session and cost evaluations insightfulness favoured games without demos slightly, while playability and project management areas had slightly higher means on games with demos. Across all evaluated areas, games with demos had lower variances than those without demos. The difference was most prominent in mechanics and playability areas, where the difference in the variances was nearly double. The means and variances are presented in figure 9.

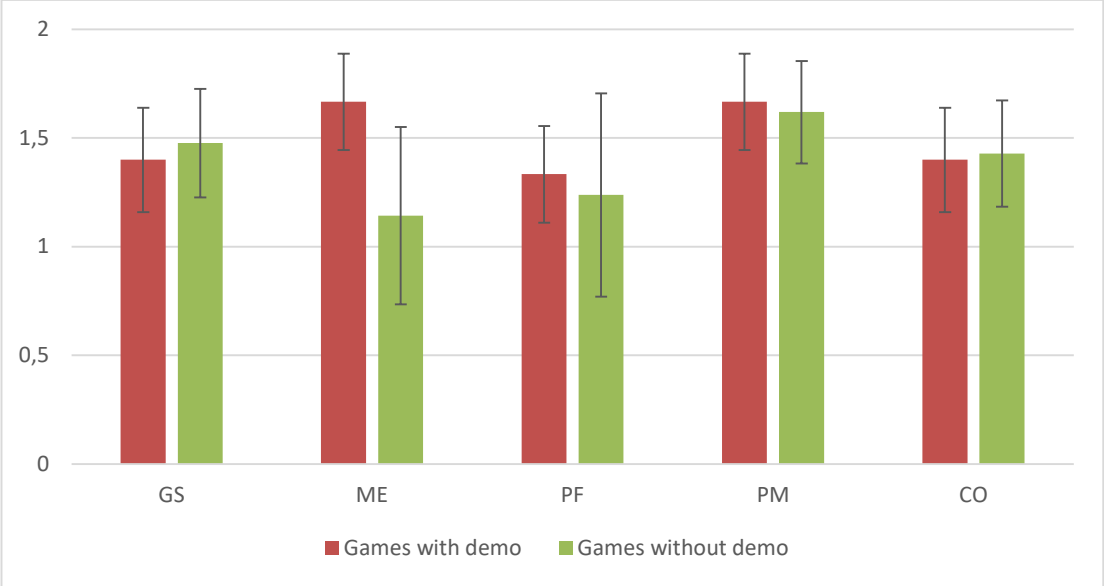

Figure 9. The mean insightfulness of each evaluation area broken down on the availability of demo.

When examining insightfulness by evaluation point, the differences between the groups on mean evaluations become clearer. For the game session evaluation area, the games without demos had higher scores in flexibility, GS1, and connection of social aspects, GS2. However, the games with demo options scored higher on the insightfulness of the connection of the game and session, GS3. The only area where the other group had a higher mean across all evaluation points was mechanics. Here the games with demo options were regularly higher than those without demo options. In the playability and flow of the game area, the means in insightfulness were close for the evaluations on clarity of controls, PF1 and the feedback, PF2. However, the difference in the UI evaluations, PF3, rose as differentiating factor between the groups. In the connection to the project management key aspects, the connection to the project management areas had the highest score. In evaluating coverage of stated learning goals, PM2, and the usage of project variables, PM3, the insightfulness switched between the groups. In the costs area, the considerations of staff and available material were rated higher in the valuation of insights provided. However, the versatility evaluation set the score neutral between the groups. These means are presented in figure 10.

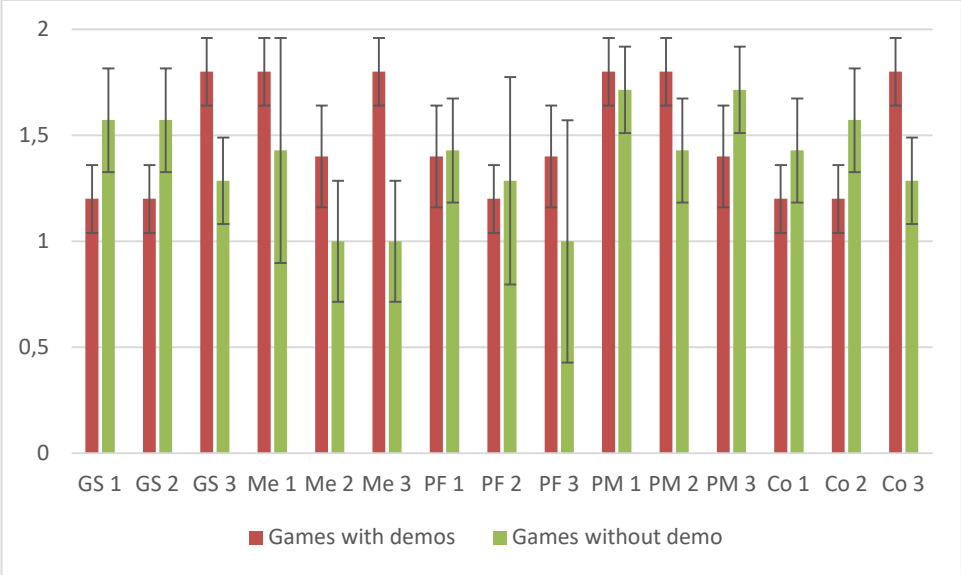

Figure 10. Means and variances of insightfulness by evaluation point.

For all but one area, the variance of insight was higher for the games without a demo option. Only on the number of project variables used in the evaluation, the variance was smaller for the games without a demo. This area also had the smallest difference between the variances of the groups. The most considerable difference in variation between the groups is in the connection of the gameplay loop and the learning goals, followed by the feedback provided by the game and the UI evaluation. The evaluation items with the highest variance on the games without demos were also these three evaluation items. The highest variance in the games with demos was how the mechanics shape the core loop, clarity of controls, the UI, and project variables' usage.

5 DISCUSSION

First, four interesting results are discussed in more detail. These rose from the results presented in the earlier section.

5.1 Highs scores on the project management evaluations

Both of the groups had their highest scores on the project management areas evaluation. This points towards it being one of the core parameters around which these games have been designed. Keeping the focus on the core learning point was one of the key principles of serious game development highlighted by the literature (Zhonggen 2019). In the evaluations, this is the only area, where no game had zero evaluation score on the evaluation points. For evaluating the connection of project management subject areas, a single game was responsible for all of the variance in the dataset. The marketing material for this game could have been for any corporate workshop training product. The page focused on the project management areas covered, and results achieved by gamified training and lacked any screenshots or mentions of the game's mechanics.

The high mean score could also point toward a too lax scoring scale from the models' point of view. Two additional evaluation points from other sections had connections to the core teaching principles of the game: the connection from the gameplay loop to the learning goals of the game and the ties from the provided feedback to the learning goals. As the evaluations on these two often fell lower than the evaluations of the project management category, it could support the too lax evaluation scale of the project management category. Alternatively, as the feedback loop during the game and the gameplay loop are not necessarily as high on the focal list of points for the developer, the lower results in these two evaluation points would direct towards lower scores on these metrics. Emmerich and Bockholt (2016) noted that their game design loop methodology could enhance the connection from the learning goals, but based on these results, it would be more needed for the mechanics of the games.

Project management was also one of the most accessible areas to evaluate from the ease of evaluation point of view, potentially due to having one of the more precise scales. This ease came from targeting the number of management areas or project variables hit and comparing the found evaluations with the listed records in the marketing material. This evaluation area was easier for the games with demo evaluation possibilities. Occasionally, it was harder to determine how major part some features played based on the marketing material alone. The evaluations of this area also scored high on the insight they provided, as a core part of the learning revolves around the core topic.

5.2 Low scoring on the mechanics of the games

On average, the games scored lowest in the mechanics' evaluation area. While the games with demo options were evaluated higher than those without demo options, this was the lowest mean score for both groups. One of the games reached an evaluation of two on each of the evaluation points, so it was not impossible. The mechanics were constructed for the selected project type in the said game, providing a great match in this area. However, this came with a trade-off in some other areas in the evaluation of the game, for example, rather long minimum sessions and limited cooperation possibilities. Here the game received zero scores as well, as the game's marketing material showed no information, which was noted in the previous section. An alternative approach could have been giving this game an empty score so it would not have impacted the end evaluations. However, this approach would have adjusted the marketing-based evaluations' positioning and not shown the game, which material ignored the area.

In the literature, the connection between the games' learning goals and the gameplay loop presented had strong emphasis (Guardiola 2016). Additionally, Mitgutsch and Alvarado (2012) listed the cohesiveness of the game systems in their evaluation model. These evaluations were partly incorporated into the model from the mechanics' evaluation area's first two items. Also, Emmerich and Bockholt (2016) noted that their game design loop method had only been fully implemented in three cases. This low implementation level could point toward the game's mechanics receiving lower emphasis than the other game elements, both during the design and marketing.

The game mechanics were also the most challenging area to evaluate. The mechanics scored lowest on each evaluation for both game groups, with some of the lowest variances. Especially the evaluation of the evolution of mechanics was challenging for the games without demos, as they scored the lowest of all of the evaluations on the challenge. Naturally, the evaluation of the gameplay loop and mechanics and their evaluation can be challenging without access to try the game hands-on. However, if the marketing material provided a good set of screenshots with access to the manual, evaluating this area was as easy as with the demo, even if it took slightly more time. Even though this was a challenging area to evaluate, it often revealed a lot about the game if there was a demo possibility. The evaluations on this subject area are often carried over in part to the evaluations on other areas, as these form core of the playing experience. Comparatively, the mechanics were rated the lowest on the games without demo possibility, though the group had higher variability. These statistics could point towards adjusting the framework for different evaluation scenarios; maybe do not go over all of the areas purely based on online marketing material.

5.3 A larger variance in the costs category

The games with demo possibilities had a more significant variance in the scores in the costs category. These were more than double the variances that the games without demo possibility received. Some of these games were free online games without optional purchases, which would explain why the games did not have several of the pointers which commercial solutions had. Additionally, the creators' main business was organizing corporate training for some of the games, which meant they had options going up to 700-800 € per player. For these events, the game's creators had to create training material. If this material has been created, why not use it in marketing and as optional material for academic customers. The ease of evaluation and insights it provided were interestingly nearly identical. This could stem from the fact that the information was instead readily available for products that had commercial versions, while the same was true for the games with only a free option.

5.4 High scores and low variances for specific evaluation areas

Some of the games received the best marks consistently on the questions related to the coverage of project management subject areas, their connection to learning goals, providing supporting material and the game sessions' connection to the learning event. This high scoring on some evaluation items was clear for the games without demo options and to a certain level for games with demos if the game was commercialized. At the same time, these evaluation areas had some of the lower variances, but the odds of scoring high on one of the areas rose the likely hood of the other areas scoring high as well.

As this was more prominent for the commercialized games, the first potential reason is marketing. These games have done their market research and concluded that these are points that have the potential to sell their games well. At the same time, these types of factors are prominent in training marketing and software sale pitches.

Interestingly the high scoring from these evaluation items did not translate to higher evaluations in the mechanics' area. While the mechanics' first evaluation was directly connected to the evaluations of the learning goals, the transferal of impact was not strong. It could also point towards some level of connection, as the two other mechanics evaluation items scored lower on average.

6 CONCLUSIONS

Serious Games and the related terms gamification and simulation-based learning are becoming more common educational tools in higher education. These games are played in a session in an educational context. The goal of this thesis was to craft an evaluation framework for serious games with project management as its' domain focus. This framework was used to evaluate serious project management games found online. At the start, the hypothesis was that it is possible to evaluate the suitableness of a serious game based on the mechanics of the game. Benchmarking was selected as the method for carrying over these evaluations as a form of action research. The research questions were set as follows:

1. What kind of features of serious games can be used to benchmark from a game component perspective?
2. How do different project management games compare to these features?

When it comes to the first research question, the games can be evaluated by their mechanics, playability and the connection to the subject the game is focused around. These evaluation areas compare the ways the games themselves are set up. The earlier evaluation frameworks from literature for serious games have typically selected an aspect from the game and focused on the said aspect of the game. For example, such literature can provide a more focused evaluation framework on playability. The evaluations on the topic at hand should always be set based on the goals the game needs to meet. There was no direct evaluation framework for project management, but literature on the area provided comparison areas for setting the evaluation items.

Additionally, it is beneficial to consider other factors than just the game's features. These were evaluated in the game session and costs categories. To see the game's fit to a learning event, it is important to consider the game's limitations for the event(s). If the game requires multiple rounds across several days, it can not be placed inside a single lecture. Additionally, if the game does not come with additional material on debriefing and setting up the sessions, the teacher has to prepare this material themselves. As such, it is not sufficient to only focus on the games' features. This insufficiency in focus also goes

against the hypothesis stated at the start, as it is not sufficient only to examine the game's mechanics.

Few trends rose when it came to evaluations of the games against these set features. On average, the games had focused strongly on project management, which resulted in a high mean score in the said evaluation area. On the other hand, the evaluations on the mechanics were considerably lower. One game notably received zero in this evaluation area as their marketing material revealed nothing on the game's mechanics and how it is played. Even though if said game is removed from the evaluations, the means for the area would still be the lowest across all of the evaluation areas. The playability and the flow of the game were between the other game aspect-related areas in their evaluation. Only one of the evaluations received zeroes in this area; this was the same game that received zero on the mechanics. As such, there were some considerations given to this evaluation area by all of the evaluated games.

If the games had been evaluated only in these three evaluation areas, the final order of the games in the evaluations would have been different. The top-scoring games would have remained in their places, but the games which received lower scores would have switched positions. This switch also supports the position that it is not only sufficient to evaluate the game features. It also goes against the hypothesis of the thesis. The resulting synthesis from this antithesis and hypothesis is following:

To sufficiently evaluate a serious game, the game should be examined from the game's content as well as the fit of the game to the learning event and the supporting material provided. The game's content can be evaluated from three points of view: the game's mechanics, playability and flow, and connection to the learning topic.

6.1 Practical implications

When evaluating the games and analysing the results from the evaluations, few recommendations rose towards the games, their development and marketing. These recommendations for the games can roughly be categorized into three groups related to marketing, development and additional provided materials. To start from the starting

point of the game's life cycle, the development. As the goal is to produce a great educational game, it is important to keep in mind the core learning goals and define them clearly in the development process. This connection was also highlighted in the literature. While mechanics of the game often had lower emphasis than the educational content, robust design on these fronts can provide a valuable edge in the competition. This focus would strengthen the benefit provided by the demo possibility. The literature highlights additional considerations for the development and comparison results in the UI (PF3). A poorly designed one can as week hinder the learning results.

Additional recommendations for the games rose from the additional materials and the related content they provided. The materials provided helped to assess the content in the games. Some of the games went as far as to provide potential lesson plans and how to set up the game for these purposes. These materials help establish the game in the context of the learning event and the factors that could be considered for the event. As this type of content is linked to the development of the game through the focus on the main learning goal.

From the games marketing perspective, a few considerations arose as well. When evaluating the games, a good set of screenshots and good instructional booklets with class examples are provided as a strong, if not more robust, the basis for evaluations than providing a demo. This material could also be presented in a video for some sections. Additionally, some of the games would have benefited if marketing materials were gathered around some theme, as the material occasionally was contradictorily or confusingly presented. For example, one of the games listed the number of available languages as three but only provided material and demo in English. One of the marketing focuses in the games was the availability of alternative modes, such as different projects or agile and traditional project options. The choice between the level of prework before the session can also provide more alternatives for scenarios and provide broader marketing possibilities.

6.2 Future research ideas

One of the development areas which rose from the evaluations is the potential retuning of the evaluation scales. The currently used scale resulted in high evaluations for the connection to the project management subject area. Additionally, rebalancing the scales might be in order for some of the other areas. Based on the evaluations on the insight provided, it could also be considered are all of the evaluations are suitable for each evaluation scenario depending on the information available. For example, if the mechanics' evaluations were different depending on the available material.

Only one of the evaluation areas is directly connected to the project management subject area. If a similar set of core management areas could be found on a different topic, the transfer of the framework to a different domain could be relatively simple. Additionally, specified variables should be discoverable to help frame the evaluation of the results. For example, a simulation game focusing on production scheduling could consider the scheduling systems direction, management of production process, quality management, staffing, resource management and adjustability. The variables followed could follow the average wait times, downtime, set-up times and the number of on-time deliveries. The evaluation could potentially move out of the industrial engineering and management study areas toward DevOps management. Using the framework in different subject areas could also provide information on the rebalancing needs of the other evaluations areas, as often, the development of serious games tends to get siloed between the different academic areas (Hallinger and Wang 2020b). Using the framework between different areas could help lower these siloes as the best practices identified in different academic areas serious games could be more widely adopted. The suggested fields fall still into the realm of engineering, so adapting the game to medical or nursing practice could provide interesting insight into the development of serious games.

Additionally, the framework does not consider the accuracy versus the realism of the game. If potentially simplifying the game could provide a more substantial learning experience, should this be given more emphasis on the evaluation model? For the earlier mentioned case, where the connection to the games was rather loose, would a third category of a seminar or workshop type service focused packages provide additional

insights. The scoring on the project management evaluation area could also be described as a more descriptive evaluation than the other area. Here the difference between one and two points mainly comes from the broadness of the evaluated game's project management covering. A game scoring one for an evaluation item could be better for a specific setting than a game scoring two. For an introductory course, a game covering only a few areas with deacent depth can provide better results than a game with a large umbrella of covered management areas, as the players can focus more on said individual learning points. As such, the game's fitting to the event is more important than the actual broadness of the project management subject areas.

As many of the games had options targeting corporate training, there should be direct considerations for this use case. In this scenario, it could be reasonable to directly reduce the emphasis on academic factors or increase the emphasis on the services provided, such as hosted workshops. The simulations could also be customized for a specific scenario where the directors are given a simulation of the current situation, and they can consider the implications of their decisions. This consideration would also set different requirements for the accuracy of the simulation.

6.3 Limitations of the research

The research goal was to construct a benchmarking framework suitable for evaluating serious games in project management. As a core part of benchmarking is incorporating the found best practices into action, this had to be limited from the thesis due to limitations on time. The validity, reliability and generalisability of this thesis are evaluated based on the framework presented by Noble and Smith (2015).

Validity is ” *The precision in which the findings accurately reflect the data*”(Noble and Smith 2015). The author has been as precise with the findings as possible. The data presented and the statistical analyses carried out on the dataset have been conducted with care. It is possible that the data on which the finding have been based can change, depending on the evaluators and the development of the games. It is also possible that a larger group could have resulted in different evaluations for the games.

Reliability is" *The consistency of the analytical procedures, including accounting for personal and research method biases that may have influenced the findings"* (Noble and Smith 2015). Through earlier research and personal and professional experiences of the researcher can impact the results, the author has been as objective as possible. A larger group of evaluators could have had different evaluation results for the framework. Preforming the sampled games' searches at different times could have impacted the formed sample, but this was reduced by incorporating two search rounds.

Generalisability is "*The transferability of the findings to other settings and applicability in other contexts*" (Noble and Smith 2015). The framework has only one connection area to the topic of the serious game. The benchmarking framework can be transferred to evaluating other serious games in other fields. As such, the framework can provide a generalisable basis for evaluations of serious games.

REFERENCES

AbleSim, 2022. *Project Management Activities* [online]. Available from: https://ablesim.com/ [Accessed 8 Mar 2022].

Adams, J., Khan, H. T., Raeside, R. and White, D. I., 2007. *Research methods for graduate business and social science students*. 1st ed. London: SAGE Publications Pvt. Ltd.

AlbaSim, 2020. *ALBASIM* [online]. Available from: https://www.albasim.ch/en/our-serious-games/ [Accessed 10 Mar 2022].

Almeida, F. and Simoes, J., 2019. The role of serious games, gamification and industry 4.0 tools in the education 4.0 paradigm. *Contemporary Educational Technology*, 10 (2), 120–136.

Anand, G. and Kodali, R., 2008. Benchmarking the benchmarking models. *Benchmarking*, 15 (3), 257–291.

Andersen, B. and Moen, R. M., 1999. Integrating benchmarking and poor quality cost measurement for assisting the quality management work. *Benchmarking: An International Journal* [online], 6 (4), 291–301. Available from: https://www.emerald.com/insight/content/doi/10.1108/14635779910288541/full/html.

Artto, K., Martinsuo, M. and Kujala, J., 2008. *Project Business*. Project business. Helsinki: WSOY.

Atweh, B., Kemmis, S. and Weeks, P., 1998. *Action research in practice : partnership for social justice in education*. 1st ed. London ; New York: Routledge.

Baba, D., Mohd Yusof, S. and Azhari, S., 2006. *A benchmarking implementation framework for automotive manufacturing SMEs*. Benchmarking: An International

Journal.

Björk, S., Lundgren, S. and Holopainen, J., 2003. Game Design Patterns. *Level Up* [online], 54 (3), 180–193. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.10.4097&rep=rep1&type=pdf.

Bronte-stewart, M., 2015. Computing and Information and Failure using a Project Status Model, 19.

Business Simulations Ltd, 2022. *Business Simulation Games & Learning Apps* [online]. Available from: https://businesssimulations.com/products/ [Accessed 10 Mar 2022].

Caccamese, A. and Bagantini, D., 2012. Beyond the iron triangle: year zero. *PM World Journal* [online], 2 (12), 1–18. Available from: http://www.pmi.org/learning/beyond-iron-triangle-year-zero-6381.

Camp, R. C., 1989. *Benchmarking : the search for industry best practices that lead to superior performance*. Milwaukee (Wis.): Quality Press.

Cesim Oy, 2022. *Project Management Simulation* [online]. Available from: https://www.cesim.com/simulations/cesim-project-management-simulation [Accessed 6 Mar 2022].

Cowley, B., Fantato, M., Jennett, C., Ruskov, M. and Ravaja, N., 2013. Learning when serious: Psychophysiological evaluation of a technology-enhanced learning game. *Educational Technology and Society*, 17 (1), 3–16.

Double Masters, 2022. *Simulations* [online]. Available from: https://www.doublemasters.com/products/ [Accessed 9 Mar 2022].

Ebbesen, J. B. and Hope, A., 2013. Re-imagining the iron triangle: embedding sustainability into project constraints. *PM World Journal*, 2 (III), 1–51.

Emmerich, K. and Bockholt, M., 2016. Serious games evaluation: Processes, models, and concepts. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9970 LNCS, 265–283.

Fong, S. W. and Cheng, E. W. L., 1998. Fong et al 1998 --- Benchmarking - A general realding for management practitioners.pdf, 407–418.

GamesByRobc, 2022. *The Project Management Game* [online]. Available from: http://thatpmgame.com/ [Accessed 6 Mar 2022].

Girard, C., Ecalle, J. and Magnan, A., 2013. Serious games as new educational tools: How effective are they? A meta-analysis of recent studies. *Journal of Computer Assisted Learning*, 29 (3), 207–219.

González Sánchez, J. L., Gil Iranzo, R. M. and Gutiérrez Vela, F. L., 2011. Enriching evaluation in video games. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6949 LNCS (PART 4), 519–522.

Guardiola, E., 2016. The gameplay loop: A player activity model for game design and analysis. *ACM International Conference Proceeding Series*.

Hallinger, P. and Wang, R., 2020a. The Evolution of Simulation-Based Learning Across the Disciplines, 1965–2018: A Science Map of the Literature. *Simulation and Gaming*, 51 (1), 9–32.

Hallinger, P. and Wang, R., 2020b. Analyzing the intellectual structure of research on simulation-based learning in management education, 1960–2019: A bibliometric review. *International Journal of Management Education* [online], 18 (3), 100418. Available from: https://doi.org/10.1016/j.ijme.2020.100418.

Hinde, D., 2012. Overview of PRINCE2. *In*: *PRINCE2 Study Guide* [online]. John Wiley & Sons, 528. Available from: https://masterworkshop.skillport.com/skillportfe/assetSummaryPage.action?assetid

=RW$9745:_ss_book:46800#summary/BOOKS/RW$9745:_ss_book:46800.

Hoenig, F., 2005. Defining computational aesthetics. *Computational Aesthetics'05 Proceedings of the First Eurographics conference on Computational Aesthetics in Graphics, Visualization and Imaging* [online], 13–18. Available from: https://diglib.eg.org/EG/DL/WS/COMPAESTH/COMPAESTH05/013-018.pdf.abstract.pdf;internal&action=action.digitallibrary.ShowPaperAbstract.

International Centre for Complex Project Management, 2022. *Online Complex Project Simulation* [online]. Available from: https://iccpm.com/complex-project-simulation/ [Accessed 7 Mar 2022].

Jetmarová, B., 2011. Benchmarking - Methods of raising company efficiency by learning from the best-in-class. *E a M: Ekonomie a Management*, 14 (1), 83–96.

Jha, K. N. and Iyer, K. C., 2007. Commitment, coordination, competence and the iron triangle. *International Journal of Project Management*, 25 (5), 527–540.

Jørgensen, K., 2010. Time for new terminology? Diegetic and non-diegetic sounds in computer games revisited. *Game Sound Technology and Player Interaction: Concepts and Developments*.

Kerschbaum, F. and Terzidis, O., 2006. Filtering for private collaborative benchmarking. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3995 LNCS, 409–422.

Kyrö, P., 2004. Benchmarking as an action research process. *Benchmarking: An International Journal* [online], 11 (1), 52–73. Available from: https://www.emerald.com/insight/content/doi/10.1108/14635770410520302/full/html.

Lameras, P., Arnab, S., Dunwell, I., Stewart, C., Clarke, S. and Petridis, P., 2017. Essential features of serious games design in higher education: Linking learning attributes to game mechanics. *British Journal of Educational Technology* [online],

48 (4), 972–994. Available from: https://onlinelibrary.wiley.com/doi/10.1111/bjet.12467.

Lehtonen, M., 2014. Evaluating megaprojects: From the 'iron triangle' to network mapping. *Evaluation*, 20 (3), 278–295.

Lim, T., Louchart, S., Suttie, N., Ritchie, J. M., Aylett, R. S., Stănescu, I. A., Roceanu, I., Martinez-Ortiz, I. and Moreno-Ger, P., 2013. *Strategies for effective digital games development and implementation*. Cases on Digital Game-Based Learning: Methods, Models, and Strategies.

Lundgren, S. and Björk, S., 2003. Game Mechanics: Describing Computer-Augmented Games in Terms of Interaction. *Proceedings of Technologies for Interactive Digital Storytelling and Entertainment (TIDSE 2003)*, 45–56.

M.Nazry, N. N. and Romano, D. M., 2017. Mood and learning in navigation-based serious games. *Computers in Human Behavior* [online], 73, 596–604. Available from: http://dx.doi.org/10.1016/j.chb.2017.03.040.

MIT Sloan School of Management, 2022. *Project Management Simulation* [online]. Available from: https://mitsloan.mit.edu/teaching-resources-library/project-management-simulation [Accessed 5 Mar 2022].

Mitgutsch, K. and Alvarado, N., 2012. Purposeful by design?: A serious game design assessment framework. *Foundations of Digital Games 2012, FDG 2012 - Conference Program*, 121–128.

Munns, A. K. and Bjeirmi, B. F., 1996. The role of project management in achieving project success. *International Journal of Project Management*, 14 (2), 81–87.

Nielsen, J., 2000. *Why You Only Need to Test with 5 Users* [online]. Nielsen Norman Group. Available from: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/.

Noble, H. and Smith, J., 2015. Issues of validity and reliability in qualitative research. *Evidence-Based Nursing*, 18 (2), 34–35.

Noemí, P.-M. and Máximo, S. H., 2014. Educational Games for Learning. *Universal Journal of Educational Research*, 2 (3), 230–238.

O'Flaherty, J. and Costabile, M., 2020. Using a science simulation-based learning tool to develop students' active learning, self-confidence and critical thinking in academic writing. *Nurse Education in Practice* [online], 47 (June), 102839. Available from: https://doi.org/10.1016/j.nepr.2020.102839.

Palmatier, R. W., Houston, M. B. and Hulland, J., 2018. Review articles: purpose, process, and structure. *Journal of the Academy of Marketing Science* [online], 46 (1), 1–5. Available from: http://link.springer.com/10.1007/s11747-017-0563-4.

Paradigm Learning, 2022. *EMPOWERING INDIVIDUALS WITH THE CORE PRINCIPLES OF PROJECT MANAGEMENT* [online]. Available from: https://paradigmlearning.com/products/countdown [Accessed 8 Mar 2022].

Pinelle, D., Wong, N. and Stach, T., 2008. Heuristic evaluation for games: Usability principles for video game design. *Conference on Human Factors in Computing Systems - Proceedings*, 1453–1462.

Pollack, J., Helm, J. and Adler, D., 2018. What is the Iron Triangle, and how has it changed? *International Journal of Managing Projects in Business*, 11 (2), 527–547.

Project Management Institute, 2017. *A Guide to the Project Management Body of Knowledge (PMBOK® Guide), Sixth Edition* [online]. A Guide to the Project Management Body of Knowledge (PMBOK® Guide), Sixth Edition. Project Management Institute. Available from: https://masterworkshop.skillport.com/skillportfe/assetSummaryPage.action?assetid=RW$647:_ss_book:132589#summary/BOOKS/RW$647:_ss_book:132589.

Regeer, B. J., Hoes, A. C., van Amstel-van Saane, M., Caron-Flinterman, F. F. and

Bunders, J. F. G., 2009. Six guiding principles for evaluating mode-2 strategies for sustainable development. *American Journal of Evaluation*, 30 (4), 515–537.

Salvador-Ullauri, L., Acosta-Vargas, P. and Luján-Mora, S., 2020. Accessibility evaluation of video games for users with cognitive disabilities. *Advances in Intelligent Systems and Computing*, 1131 AISC, 853–859.

Shenhar, A. J., 2004. Strategic Project Leadership®: Toward a strategic approach to project management. *R and D Management*, 34 (5), 569–578.

Simulation Powered Learning, 2021. *PROJECT MANAGEMENT AND AGILE SIMULATIONS* [online]. Available from: https://simulationpl.com/ [Accessed 6 Mar 2022].

Snyder, H., 2019. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research* [online], 104, 333–339. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0148296319304564.

STS Sauter Training & Simulation SA, 2022. *SimulTrain* [online]. Available from: https://sts.ch/en/products/simulation/simultrain [Accessed 5 Mar 2022].

Sweetser, P., Johnson, D. and Wyeth, P., 2012. Revisiting the GameFlow Model with Detailed Heuristics. *Journal of Creative Technologies* [online], 2012 (3), 1–16. Available from: http://journal.colab.org.nz/article/21.

Sweetser, P. and Wyeth, P., 2005. GameFlow. *Computers in Entertainment* [online], 3 (3), 3–3. Available from: https://dl.acm.org/doi/10.1145/1077246.1077253.

The Crowd Training, 2018. *Freely Explore Some of What PM City Has to Offer* [online]. Available from: https://www.thecrowdtraining.com/free-content [Accessed 10 Mar 2022].

Zairi, M. and Leonard, P., 1994. *Practical benchmarking : the complete guide*. London: Chapman & Hall.

Zhonggen, Y., 2019. A Meta-Analysis of Use of Serious Games in Education over a Decade. *International Journal of Computer Games Technology*, 2019 (3).

Appendix 1. Formated and condensed evaluation form

Game session	GS 1. How long is the session/sessions
	GS 2. How do social aspects tie to the session
	GS 3. How the game ties into the learning event
Mechanics	Me 1. How gameplay loop connects to learning goals
	Me 2. How the mechanics connect to the core gameplay loop
	Me 3. How mechanics and loop and their impact evolve as the game progresses
Playability and flow of the game	PF 1. Clarity of controls and goals in-game
	PF 2. How feedback provided during the session ties to learning goals and the clarity of feedback
	PF 3. Do the UI and aesthetic choices support learning
Coverage of key aspects of project management	PM 1. Which of the management areas of project management are covered based on the listing in section 2.4.3
	PM 2. How do the selected management areas cover the stated learning goals of the game
	PM 3. How many of the project variables were used in-game
Costs	Co 1. The time required from staff
	Co 2. The materials that are provided in the package
	Co 3. How versatile are the project setup options

Appendix 2. Evaluation scores of the games.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
GS 1	1	1	2	1	1	1	1	1	1	2	1	1
GS 2	1	2	2	2	0	2	2	1	1	2	0	1
GS 3	2	2	2	2	1	2	2	2	1	1	2	2
Me 1	2	2	1	1	1	1	2	0	2	2	1	1
Me 2	2	2	1	2	1	1	1	0	1	1	1	1
Me 3	2	1	1	1	1	1	1	0	1	1	1	0
PF 1	2	2	1	1	2	1	2	1	2	1	1	1
PF 2	1	2	1	1	1	1	2	0	1	2	1	1
PF 3	2	1	2	2	1	2	2	0	2	1	1	1
PM 1	2	2	2	2	1	1	2	2	1	2	2	2
PM 2	2	2	2	2	1	2	2	2	2	1	2	2
PM 3	2	2	1	2	1	1	1	2	2	2	2	2
Co 1	2	1	2	2	2	2	2	1	1	0	2	2
Co 2												

Co 3	1	2	2	2	1	2	2	2	2	0	2	2
Co 4	2	1	2	1	0	1	1	1	2	2	1	1