University of Twente
Faculty of Behavioral, Management and Social Sciences (BMS)
Department of Human Factors & Ergonomics
Bachelor's thesis

Can Mobile Gaming Habits Predict the Perceived Usability of Educational Games?

Christof Schulz

Supervision and Examination:

Dr. Marleen Groenier

Dr. Simone Borsci

August 2020

Foreword

This thesis was written under special circumstances caused by the worldwide outbreak of Covid-19 and internal issues in the supervision of the students. The examination board of the University of Twente granted an extension of the deadline for the submission of the thesis. Due to the unexpected setbacks, the study was conducted differently than we planned initially. A data collection on real simulators was not possible as the university buildings were locked for common use and working with simulated data was not possible due to organizational issues. Therefore, a usability assessment of two applications served as a quick replacement for a topic of the thesis. As it was not possible to conduct the usability study in a laboratory setting either, it had to be conducted online. This thesis was started at the 02/06/2020 and finalized at the 23/08/2020.

Abstract

Background: Minimally invasive surgery (MIS) is a common surgery practice nowadays. MIS entails several advantages compared to open surgery but learning MIS can be challenging due to the unique characteristics of the procedure. Technological advances made it possible to simulate MIS so surgeon apprentices can practice in a safe environment. As simulator training can be related to barriers, mobile applications could be used as additions to existing training methods. To assess the satisfaction with available applications, the perceived usability can be assessed. Two applications, SimuSurg and TouchSurgery, were described and assessed in terms of perceived usability. Isbister and Schaffer (2008) described that being a frequent gamer can influence the perceived usability of a game. This study aims to show the influence of frequent gaming on the perceived usability of the two applications.

Methods: A within-subject design was employed. Participants were recruited through convenience sampling and were asked to do the study remotely and online. A specific subset of tasks related to MIS was performed by the participants in each of the applications. After each task, the System Usability Scale (SUS) was utilized to measure the perceived usability for each application. Participants were divided into three categories of gaming frequency. The reliability analysis of the SUS with Cronbach's was found as good.

Results: TouchSurgery received higher SUS scores than SimuSurg overall. Mobile gaming frequency did not have an influence on the perceived usability of TouchSurgery and SimuSurg. Gender did not have a significant effect on the perceived usability of TouchSurgery and SimuSurg, either.

Discussion: Results suggest that being a gamer or having more gaming experience does not influence the perceived usability of game-like applications. The claim of Isbister and Schaffer (2008) has not been researched in scientific literature to the best of our knowledge and needs further investigation. The results of the study should be interpreted carefully, as it had several limitations such as a high dropout rate, no laboratory conditions and inaccurate gaming frequency groups.

Table of contents

1. Introduction

1.1 Background

Alternative methods of surgery became more popular in the last decades. One of these approaches is called minimally invasive surgery (MIS), a practice in which the surgeon operates through small incisions in the patient's skin to gain access to the body. Through one of these incisions the laparoscope is navigated into the body and the remaining ports allow a variety of specially designed surgical instruments to enter the operating area (Soper et al., 1994; Frecker et al., 2007).

However, this technique requires the surgeon to have a different set of skills compared to open surgery. MIS has a different learning curve than open surgery because it introduced a new and unique set of psychomotor skills for a surgeon to master (Gallagher & Smith, 2003; Spruit & Band, 2014). Traditionally surgery is trained in the operating room with the apprenticeship model (Basdogan, Sedef, Harders, & Wesarg, 2007). The surgeon apprentice observes, assists and imitates the procedure from an experienced surgeon (Spruit & Band,2014; Yiannakopoulou, 2015). In the recent years, this learning model was examined and questioned in terms of its efficiency. A larger number of people is dying from medical mistakes each year, also caused by the inexperience of operating surgeons (Basdogan et al., 2007). The increasing complexity of procedures and devices, a reduction of training hours in Europe and North America, and concerns regarding patient safety require novel training methods next to the apprenticeship model (Sugand, Mawkin, & Gupte, 2015). Especially for MIS alternative training methods are required because the unique skills needed for MIS are more difficult to learn by observation (Stefanidis et al., 2005).

Consequently, simulation became an important component of the education of surgeons. Technological advances made it possible to simulate MIS on different interfaces. Training on box trainers, video trainers and advanced simulators became a regular component in many curricula (Alaker, Wynn, & Arulampalam, 2016). Studies showed that this type of training leads to better performance and more resistance to skill decay compared to traditional learning without simulators (Gurusamy, Aggarwal, Palanivelu, & Davidson, 2008; Stefanidis, Korndorffer, Markley, Sierra, & Scott, 2006). Moreover, simulations can recreate scenarios that are rarely encountered in reality to prepare professionals for challenging situations. After each training session, the trainee can get extensive feedback on his/her performance from the system (Aggarwal et al., 2010).

Despite the obvious benefits of simulator training, there are also barriers. High quality simulators are extremely expensive. A fully loaded high-fidelity simulator system costs at least 60.000€ (surgicalscience, n.d.) and can therefore only be afforded by some educational institutions. Furthermore, the simulator is bound to one place, and it can only be used if the location of the simulator is accessible. Even two weeks without practice on a simulator can lead to a substantial decline in skills (Kerfoot & Kissane, 2014). Although mechanical and VR constructs allow surgeons to acquire psychomotor and technical skills, they neglect an emphasis on important non-technical skills like cognitive decision-making. Decision-making contributes to demonstration of clinical and non-technical competency (Franko & Tirrell, 2012). If students want to train non-practical skills like decision making and task rehearsal, a more accessible training device, like for instance a mobile phone could be helpful (Sugand, Mawkin & Gupte, 2015). A study conducted in 2011 demonstrated that 91.8% of surgery residents use a smartphone for medical purposes. Examples for the type of applications used by the residents are drug guides, medical calculators, or textbook-like applications. The same study showed that medical personnel believes that there are too little high-quality apps available. At the same time, high-quality applications are strongly desired (Franko & Tirrell, 2012). In the course of this paper, two free of charge mobile applications that could prove useful in developing skills for MIS were assessed. To assess the quality of an app, the perceived usability should be assessed to gain more into insights the perceived efficiency, effectiveness and user satisfaction of the two applications. The names of the applications are TouchSurgery, developed by *Digital Surgery Ltd.* and SimuSurg, created by the *Royal Australasian College of Surgeons.*

1.2 Advantages, challenges and training of MIS

MIS has clear advantages for the patient compared to open surgery. Small incisions lead to smaller wounds. This causes faster healing and less postoperative pain which makes the procedure a more comfortable alternative for the patient. Secondly, the small incisions created through MIS do not create a postoperative scar, so it has cosmetic benefits. Finally, the reduced healing time leads to a faster discharge from the hospital which makes MIS more efficient for the healthcare system from an economical perspective (Novitsky et al., 2004; Jaffray, 2005).

Although it is undisputed that MIS is beneficial for the patient, it can be challenging for surgeons. Since the procedure differs from open surgery, the surgeon must acquire another set of skills to perform MIS. One of the major challenges relates to the imagery provided by the laparoscope. In laparoscopic surgery, the environment captured by the camera is displayed on a screen. The 3D information is transformed into a 2D image from a single perspective (Gallagher & Smith, 2003). This can lead to perceptual and spatial difficulties, as the depth is harder to estimate and hand-eye coordination becomes more difficult (Gallagher & Smith, 2003; Groenier, Schraagen, Miedema, & Broeders, 2014).

The movement of the surgical tools can be another issue during MIS, as the surgeon needs to take a position that can feel unnatural or uncomfortable. Furthermore, the surgical instruments are fixed to an axis, causing a "fulcrum-effect". Consequently, movement of the instrument handle from the outside leads to movement into the opposite direction inside the patient's body (Jordan, Gallagher, McGuigan, McGlade, & McClure, 2000). On top of that, surgeons receive little haptic feedback because they interact with internal organs through surgical instruments in thin, long tubes (Basdogan et al., 2007).

Gallagher and Smith (2003) emphasize that successful performance in MIS requires long-term practice and multiple training sessions. Technological innovations such as virtual reality simulation and e-learning applications show consistent improvement of learning outcomes of trainees and already play a role in surgical training programs (Graafland, Schraagen, & Schijven, 2012). One of the main barriers of simulator training is resource-intensiveness (Wang, DeMaria Jr, Goldberg, & Katz, 2016). They are expensive (Gallagher & Smith, 2003) and constrained by work-shifts and opening times (Buttussi et al., 2013). Furthermore, simulator training depends on educators that need to guide the training course in an enthusiastic and appealing way in order to be effective (Dieckmann, Friis, Lippert, & Østergaard, 2012). The level of difficulty can be too high or too low for some of the participants of the course (Dieckmann et al., 2012).

A concept that is not affected by those barriers is interactive learning through mobile applications (Graafland et al., 2012). While mobile applications cannot simulate surgical processes as precisely as simulators on one hand, they bear advantages for the trainee on the other hand. One of the advantages is easy accessibility. Almost everybody possesses a smartphone nowadays. It is constantly available for trainees to practice whenever, wherever they want (Franko & Tirrell, 2012). Trainees are not constrained by a schedule that hinders them from practicing according to their time preferences. Furthermore, training on mobile devices can be individualized and provides an autodidactic way of learning independent of educators. Previous studies have shown that the usage and demand of educational applications is high among health care workers which hints that they are perceived as useful tools (Franko & Tirrell, 2012).

1.3 Serious games and gamification – commonalities and differences

The use of educational games as learning tools is a promising approach. They can reinforce knowledge, problem-solving skills, collaboration and communication (Dicheva, Dichev, Agre, & Angelova, 2015) . A major advantage of educational games is that they have a great motivational power, as they utilize different mechanisms to encourage people to engage with them (Dicheva et al., 2015).

Although the usefulness of educational games seems evident, they can be designed quite differently. In the literature, distinctions between serious games and gamifications are made. Serious games can be defined as fully designed games in which education is the primary goal (Landers, 2014). Gamifications are defined as the use of singular elements commonly associated with video games in non-game contexts (Deterding, Sicart, Nacke, O'Hara & Dixon, 2011; Sailer, Hense, Mayr, & Mandl, 2017). Landers (2014) elaborates that there are commonalities, but also differences between gamification and serious games: "Games and gamification are similar in that they both incorporate game elements; they differ in that games incorporate a mixture of all game elements, whereas gamification involves the identification, extraction, and application process of individual game elements." Gamification interventions involve some game elements with a utilitarian purpose while serious games are designed as full-fledged games for a purpose other than just entertainment (Gentry et al., 2019). In a study conducted by Gentry et al. (2019) the authors analyzed a total of 30 gamification and serious gaming interventions. The authors stated that both gamification and serious gaming have been at least as effective as the control group, but there was insufficient evidence to state whether gamification or serious games are more effective than any other.

To create a taxonomy of all relevant attributes that constitute a serious game, Bedwell et al. (2012) organized a group of game developers and a group of gamers. Their task was to use card sorting to form categories that define a game. Bedwell et al. (2012) argued that although there is a lot of research about the use of games in educational contexts, no clear definitions have been established what elements define a game. That makes comparisons between different applications more difficult, as some of them may be categorized as games while others may not. Bedwell et al. (2012) identified following characteristics typical for a game: Action language, assessment, conflict/challenge, control, environment, game fiction, immersion, and rules/goals. The descriptions of these game elements are shown in Table 1. To categorize the applications that are about to be assessed in this paper, the game elements included in TouchSurgery and SimuSurg were compared (see Table 1).

Table 1.

Categorization of TouchSurgery and SimuSurg based on Bedwell et al. (2012)

Game element	**Description**	**TouchSurgery**	**SimuSurg**
Action language	Communication between player and system	Information about procedures and instructions for the user	Objectives of the level and instructions how to use tools
Assessment	Feedback concerning the player's objectives	Learning progress is tracked and can be assessed in a quiz	Time needed to solve the level is stopped and the player is rated based on performance
Conflict/Challenge	Presentation of problems that can be difficult to solve	No real challenges, quiz is rather rudimentary	Levels increase in difficulty and need to be solved in order to proceed
Control	Player's control over the game elements	No control within the chosen procedure, user is guided without any freedom in his/her actions	The player has full control over instruments and can influence the game elements freely
Environment	Player is immersed into physical surroundings	Simulates realistic environments like operating rooms and human bodies from the inside	Simulates a virtual space to interact with objects
Game Fiction	Elements that are disparate from the real world	No fiction	User interacts with abstract, unreal objects

Human interaction	Human-to-human contact	No human interaction	No human interaction
Immersion	The game includes perceptual and affective elements	No sound, visual stimuli are barely used	Sounds give auditory feedback, extensive use of visual stimuli (e.g. through blinking)
Rules/goals	Game has clear goals, rules and information on progress towards goals	Learning phase defines learning objectives, rules/goals are obvious	Goals are clear, some rules are written explicitly, others must be learned by trial-and-error
Conclusion		**Gamification**	**Serious game**

Based on the taxonomy of Bedwell et al. (2012), we concluded that TouchSurgery and SimuSurg resemble two different types of application. TouchSurgery can be defined as gamification as it only incorporates some of the elements of a game. The user cannot interact with the artificial environment, there are no sounds, no difficulties or challenges and scoring system. It rather resembles a video trainer in which the trainee sees a video and then tries to memorize the observations. There are only few singular game elements like swiping or dragging to fulfill certain tasks. Another game-like element is quizzing. When the app is opened for the first time, TouchSurgery is advertised with a slogan “Training, not gaming”. This underlines that it was not the designer’s intention to create a mobile game.

SimuSurg was categorized as a serious game. It incorporates all gaming elements except human to human interaction to some degree. In SimuSurg, a whole simulated environment is created. How the objective is reached is up to the user, hence the control and the user’s freedom is much higher than in TouchSurgery. SimuSurg utilizes a timer as challenge in which the player needs to beat a level, a high score is saved, and the player is rated based on his/her performance. The higher the level, the more difficult it becomes. Sound effects, flashing lights and an abstract 3D environment create a feeling of immersion. SimuSurg is designed like one of innumerable mobile games in the app stores, but the purpose and the goal of the game differ from a normal mobile game.

1.4 Could being a gamer have an influence on perceived usability?

A major challenge related to the design of serious games is that developers must design a game that is not only usable for a specific target group – e.g. hardcore gamers – but for everyone. Olsen, Procci and Bowers (2011) advised to ask for background information as it could influence the perceived usability of a game. This could involve age, gender, and basic demographic information as well as any further background information that may influence the interaction with the software. This additional information includes previous knowledge and experience with the material to be covered, gaming experience, reading level, and other user capabilities and limitations such as disabilities that may impact the interaction with the game (Olsen, Procci, & Bowers, 2011).

One of the factors that should be considered according to Olsen, Procci and Bowers (2011) is the gaming experience of the user. Casual gamers are a specific target group that has different characteristics from hardcore gamers. As Isbister and Schaffer (2008) described, casual gamers lack a set of trained skills. They have less prior experience with general conventions in games, and they have a different level of tolerance for frustration. Since failure is such an integral part of many digital games, these games train regular gamers who accept and enjoy the challenge of threat of loss. This creates a kind of player who is interested in experimentation and more patient through episodes of difficulty. Furthermore, they are more willing to struggle with a clumsy control scheme or unclear interface (Isbister & Schaffer, 2008). Casual gamers do not have those attributes, and therefore have a different set of expectations for a game. This different set of expectations leads to a different set of design criteria for a usable product. The game should be significantly less complex and punishing, and it should be more easy to use. The game would still be a learning environment, but the learning methodology would be less open–minded and more guided (Isbister & Schaffer, 2008). The combination of those factors would lead to a shorter, simpler game that is more accessible to people who have not played digital games before.

1.5 Mobile usability testing with the System usability scale

Usability testing is an evaluation method used to measure how well users can use a specific software system (Zhang & Adipat, 2005). It is a key step in the successful design of new technologies and tools, ensuring that people will be able to interact easily with innovative technologies (Moreno-Ger, Torrente, Hsieh, & Lester, 2012). The testing can help the researchers to improve the design of a certain product (Hussain, Mutalib, & Zaino, 2014). One of the biggest issues with educational games is the inadequate integration of educational and game design principles (Arnab et al., 2015). This is caused by the fact that digital game designers and educational experts usually do not share a common vocabulary and view of the domain (Arnab et al., 2015). For the design of mobile applications, the specific characteristics of mobile phones must be considered: Small screen sizes, non-traditional input methods and navigation difficulties can cause issues. Therefore, usability is a more delicate issue for mobile technology than for other areas because many mobile applications remain difficult to use or inflexible (Coursaris & Kim, 2011).

One popular way to assess the perceived usability is the system usability scale (SUS). The SUS was developed by Brooke (1996) as a “quick and dirty” way to assess the perceived usability of any system. The survey consists of ten items, and aims to capture the effectiveness, efficiency and satisfaction of a product which constitutes usability according to Brooke (1996). The survey produces a single score from 0 to 100 that can easily be understood by a wide range of people. The SUS can be considered as a flexible tool that can be used to measure usability of a wide range of interface technologies (Bangor, Kortum, & Miller, 2008). For example, it has been used to evaluate safety signs, voting systems, medical systems, computing systems, websites and mobile applications (Kortum & Sorber, 2015).

1.6 Research Questions

First, overall SUS scores of TouchSurgery and SimuSurg will be calculated and compared to see which of both applications is more favored in general. The main goal of this paper is to explore if users who play mobile games more frequently have a different perception of the usability of the applications than users who do not play mobile games frequently. Based on the Bedwell taxonomy, we concluded that SimuSurg is more game-like than TouchSurgery. Isbister and Schaffer (2008) elaborated that non-gamers prefer applications that are easy to use and more guided. Therefore, we will investigate if this assumption also transfers to the context of serious games and gamifications in medical education. As gender has shown to have an effect on perceived usability in some research contexts (Lin & Chen, 2013; Moss, Gunn, & Heller, 2006), we will also investigate whether females rate either of the applications differently from men on the SUS. Therefore, the following research questions will be investigated in the course of this paper:

1. How is the average SUS score of TouchSurgery and SimuSurg?

2. Does gaming frequency have an influence on the SUS scores of the two applications?

3. Does gender have an influence on the SUS scores of the two applications?

2. Methods

2.1 Study Design

To gather the data for the usability assessment of SimuSurg and TouchSurgery, a quantitative within-subject design was chosen. Participants were required to do the study remotely. A survey with randomized order including a usability test for each application was sent out to participants of different target groups. The whole study was conducted online, as the worldwide outbreak of SARS-CoV-2 disallowed us to conduct a study which required physical presence. This study was approved by the ethics committee of the University of Twente (request no. 200884).

2.2 Participants

Participants for this study were recruited via convenience sampling. We contacted three different target groups: The first group was labelled *novices* and it consisted of active students without any medical background. The second group was called *intermediates,* medical students that had at least one year of studying in their domain. We set at least one year of studying as a requirement to ensure that the students had at least some familiarities with the standard procedures in medicine. The third group, *experts*, consisted of surgeons and nurses with at least two years of job experience. It was assumed that experts with at least two years of work experience would be familiar with the procedures shown in the tasks. As the participants were supposed to test mobile applications, it was required for them to possess a smartphone. To fill out the survey simultaneously, they also needed to have a computer or a laptop available. Participants who had prior experience with the apps needed to be taken out to ensure that none of the participants were biased due to pre-existing beliefs. It was also necessary for them to fill out the whole survey in a minimum amount of 20 minutes, as it can be assumed that it is not manageable to fill out the survey in less time properly.

A total of 78 participants of different expertise opened the survey (figure 1). Since many of them opened the survey without completing it, 26 blank entries were removed. 19 participants filled out the demographics only and stopped afterwards, so we removed them from the dataset as well. Another 2 had to be taken out due to prior experience with TouchSurgery. One of the participants encountered heavy technical issues with TouchSurgery and could not perform the tasks assigned, so he was taken out. One of the participants needed less than ten minutes for the whole study. We took him out as we believed that it is impossible to do the tasks in such a short period of time. It should be noted that 4 of the participants we included either forgot to upload a screenshot or uploaded a wrong screenshot. However, they completed the rest of the study in a reasonable way. We decided to include these participants, as a screenshot can be easily forgotten.

After excluding 48 out of 78 participants, we were left with a total of 30 valid responses. Participants ranged in age from 19 to 49 (M=25.63). We had 16 male and 14 female respondents. 17 of the participants were part of the *novice* group, 6 were *intermediates,* and 7 were *experts.* 11 participants indicated they *never* play mobile games. 3 participants responded that they play mobile games 1-3 times per month which was labelled *rarely.* 9 participants responded to play mobile games 1-3 times per week which was labelled *sometimes.* 7 participants stated to play mobile games more than 3 times per week, which made them part of a group that plays *often.*

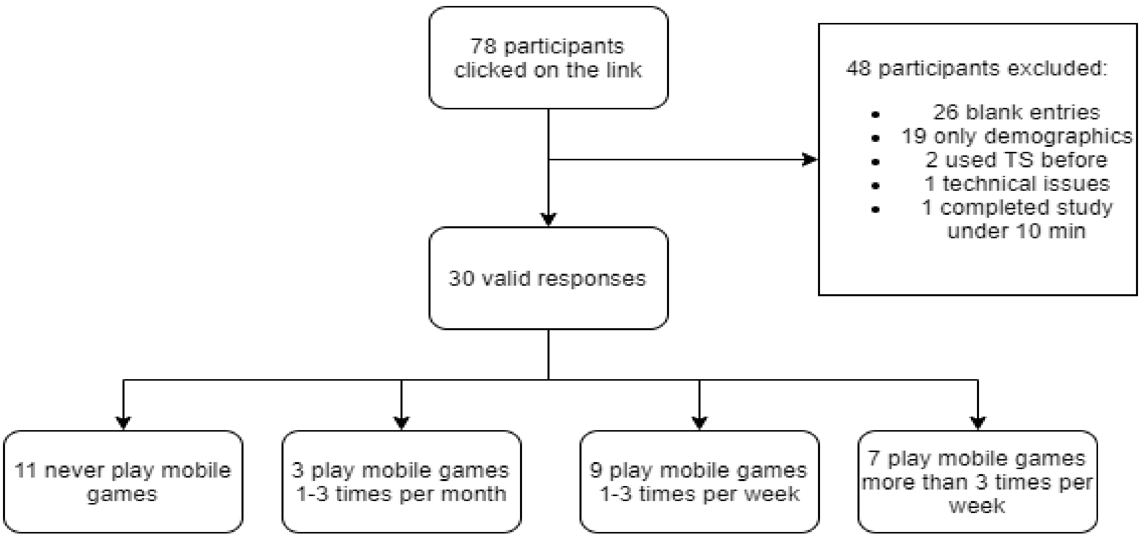

Figure 1. Flowchart of the participants with exclusion criteria and grouping of the responses based on mobile gaming frequency.

2.3 Materials

Letter of invitation, information brochure, informed consent, Qualtrics

To recruit participants for the study, we sent a letter of invitation (Appendix A) and an information brochure (Appendix B). An informed consent with six statements was also included (Appendix C). We used Qualtrics to create surveys, the information brochure and all relevant instructions. The survey software was chosen because it allowed us to randomize the order of the survey and access the data easily.

Devices

To take part in the online study, the participants had to use their own smartphones as well as their own personal computers. This was necessary, as participants were asked to install and run the applications on their phones while filling out the survey on the computer at the same time. A standardization of the devices was not possible, as the study was conducted online due to the limitations caused by COVID-19.

Demographics questionnaire

To gather data about the background of the participants, we asked them to fill out a short survey about demographics. It included information like age, gender, and level of medical education. Furthermore, the smartphone model used to run the applications and frequency of mobile/video gaming were asked for.

Instructions

To ensure that the correct applications were used in the intended way, we included detailed step-by-step instructions to demonstrate how to find, download and set-up both applications. Additionally, we added instructions on how to take and upload a screenshot as a proof of completion. Finally, a guide how to delete the TouchSurgery account was added. For most of the instructions we used images to make the process as clear as possible. For this study, we needed to write very precise instructions as participants had to fulfill the task remotely without being able to ask any of the researchers for help immediately. A small pilot test showed that the instructions had to be as detailed and unambiguous as possible, as the participants of the pilot study sometimes did not understand what exactly they were supposed to do.

System Usability Scale

For the usability test, the system usability scale (SUS) was used. The SUS is a quick and easy way to assess usability and was originally developed by Brooke (1996). It is a measure that consists of 10 items with a 5-point Likert Scale to assess the perceived usability of each of the applications (Appendix D). A minor adjustment was made, as the term “system” in the scale was replaced with “application” to avoid ambiguity. In order to calculate the participant’s final score, the rating of each item is rated from 0 to 4, aggregated and multiplied with 2.5. Final scorings can range from 0 to 100.

To evaluate the scorings on the SUS, benchmarks have been defined to constitute a grading system for usability (Lewis & Sauro, 2018). The grading scheme is shown in Table 2.

Table 2.

Grading of the SUS introduced by Lewis & Sauro (2018)

Grade	SUS	Percentile range
A+	96-100	84.1 - 100
A	80.8 - 84	90 - 95
A-	78.9 - 80.7	85 - 89
B+	77.2 - 78.8	80 - 84
B	74.1 - 77.1	70 - 79
B-	72.6 - 74.0	65 - 69
C+	71.1 – 72.5	60 - 64
C	65.0 – 71.0	41 - 59
C-	62.7 – 64.9	35 - 40
D	51.7 – 62.6	15 - 34
F	0 – 51.6	0 - 14

Net Promotor Score

The Net Promotor Score (NPS) is a tool to assess how satisfied the participant is by assessing how likely the participant will recommend the product to another person (Hamilton et al., 2014). The score is measured through a singular question: “How likely are you to recommend the product to a friend/colleague?”, which is rated on a 10-point Likert scale. Users who choose a scoring from 1 to 6 are labelled detractors and will actively discourage others to use the application. Scores 7 and 8 are passives who are broadly happy with the app, but would not actively promote it. Scores 9 and 10 are promoters, who would definitely recommend and use the service again (Hamilton et al., 2014).

Post-task questionnaire

After each of the two tasks, a post-task questionnaire was shown. It consisted of three questions. Participants were asked if they used the application before, if they successfully completed the task and if they encountered any technical issues. Furthermore, they could add additional remarks and comments.

End of survey questionnaire

At the end of the study, participants could indicate which application they liked better. Options were TouchSurgery, SimuSurg, neither or both. We implemented a comment box for final remarks.

2.4 Description of the apps and tasks

The exact task descriptions handed out to the participants can be viewed in Appendix E.

SimuSurg

SimuSurg is an application that simulates a MIS setting and provides a novel way to learn about the fundamental aspects of surgical skills. While it is obvious that SimuSurg cannot replace high-fi simulator training due to various limitations caused by the properties of a mobile phone, it can at least provide a basic understanding of how MIS works. The imagery from the insides of the body is displayed on a screen which is similar to the camera perspective in MIS. The instruments used in the game resemble some of the instruments used in real procedures. Examples are forceps, graspers and specimen bags (Reddy, 1994). Although many elements are different from real simulator training, SimuSurg can be used to become familiar with some of the basic mechanics of laparoscopic surgery.

In the SimuSurg app, participants had to complete a task consisting of 12 levels. After installing and starting SimuSurg from the home menu, participants had to press “Start”. Then, “Beginner” had to be selected and the first level “Scope introduction” was started. After an unsuccessful completion of a level, users could simply try again an indefinite amount of times. After completing a level, they could proceed to the following level by hitting “Next activity”. After the participants managed to complete the 12th level “irrigation introduction”, they were asked to take a screenshot and the task in SimuSurg was completed. If participants did not manage to reach the 12th level within 30 minutes, we advised to give up and proceed with the survey without trying further.

Google Playstore link: https://play.google.com/store/apps/details?id=com.cmee4.endoapp&hl=de

Apple AppStore link: https://apps.apple.com/de/app/simusurg/id1174517345

TouchSurgery

TouchSurgery is a learning platform in which various surgical procedures can be observed and rehearsed. TouchSurgery lacks many elements that games usually have, and it rather resembles an internet database on which people can educate themselves. It includes more than 200 simulations and videos in the library. The user can learn the steps, instruments and anatomy involved in surgical procedures. TouchSurgery has been confirmed as a useful tool for surgical education in various studies. The content, face and construct validity has been confirmed (Sugand, Mawkin, & Gupte, 2015). Furthermore, the authors stated that the application is an effective adjunct to traditional learning methods and has potential for curriculum implementation. Bunogerane et al. (2018) concluded that TouchSurgery is a useful tool to improve knowledge and technical skills but has certain limitations in the type of experience that can be gained. They suggested TouchSurgery as a tool that could mostly be used in low- to middle-income countries.

The task our participants were supposed to fulfill in TouchSurgery was called "laparoscopic appendectomy". It resembles a typical minimally invasive procedure (Long et al., 2001). After installing the app and setting up an account, the users entered the main menu. To find the respective surgery, the users had to tap the magnifying glass which illustrates the search function. Next, "laparoscopic appendectomy" was supposed to be typed into the search term box. Then, the first result had to be chosen. In TouchSurgery, the laparoscopic appendectomy consists of three learning phases and training phases. Our participants only had to do the learning phases. After the participants completed the third learning phase, they were supposed to take a screenshot and the TouchSurgery phase was finished successfully.

Google Playstore link: https://play.google.com/store/apps/details?id=com.touchsurgery&hl=de

Apple AppStore link: https://apps.apple.com/de/app/touch-surgery-surgical-videos/id509740792

2.5 Procedure

The procedure of the study is depicted in figure 1. Participants received a letter of invitation with a link to the whole questionnaire. The link guided them to Qualtrics. Firstly, the information brochure appeared. When clicking next, participants were supposed to read the informed consent. To continue with the study, they had to agree with six statements. Afterwards, the demographics questionnaire had to be filled out. Next, the instructions on how to set up each of the apps were provided. After the participants continued, they encountered the first task.

To ensure that the order of the tasks did not influence the participants, we randomized the tasks via Qualtrics. Therefore, participants either had to do the SimuSurg or the TouchSurgery task first. Consequently, the description of either the SimuSurg or the TouchSurgery task appeared. For the TouchSurgery task, the participants had to create an account before being able to use the application. Detailed instructions on how to create the account were provided. After finishing a task, we included instructions that showed how to take a screenshot. Then, we asked the participants to fill out the SUS, NPS and the post-task questionnaire. If participants indicated that they encountered difficulties that disallowed them to complete the task, then the SUS and the NPS were skipped and they were directed to the post-task questionnaire right away.

Afterwards, they proceeded to the next task. Now the description of the second application was shown. Upon completion, SUS, NPS and post-task questionnaire had to be filled out. Once again, the SUS and NPS were skipped if the participants encountered difficulties. After finishing both tasks, the participants were asked to upload one screenshot for each task and to fill out a small post-survey questionnaire. Before the end of the study, we provided a quick guide on how to uninstall both apps and how to delete the TouchSurgery account.

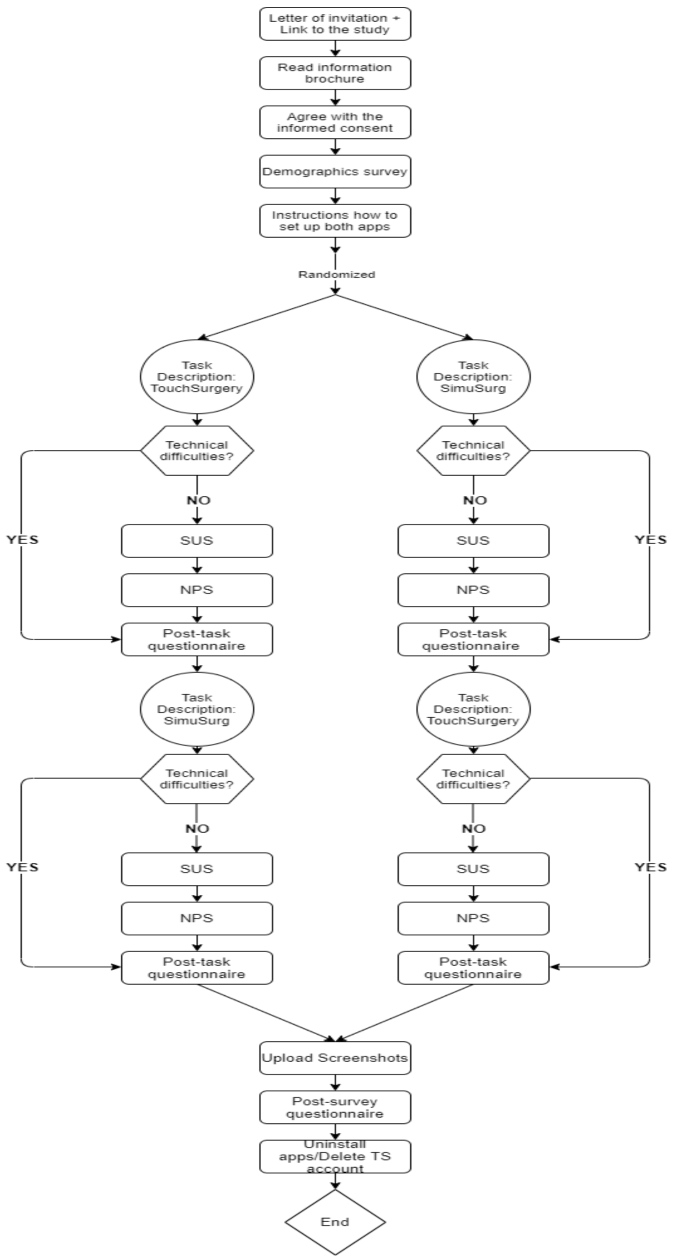

Figure 2. Flowchart of the procedure

2.6 Data processing

Reforming mobile gaming groups

As the *rarely* mobile gaming category only consisted of 2 participants, it was questionable if such a small sample is suitable for statistical analyses. It was decided to include them into the *sometimes* group, creating a combined category with 12 members that played once per month to three times a week. Concludingly, we were left with three different mobile gaming frequency groups: *Never, sometimes* and *often.*

Outlier analysis

To see if there are outliers that could influence the results, a quick outlier analysis with boxplots was performed. In the boxplot of SimuSurg, it can be observed that respondent no.6 is an outlier and may have a strong impact to sensitive statistical tests among a small population.

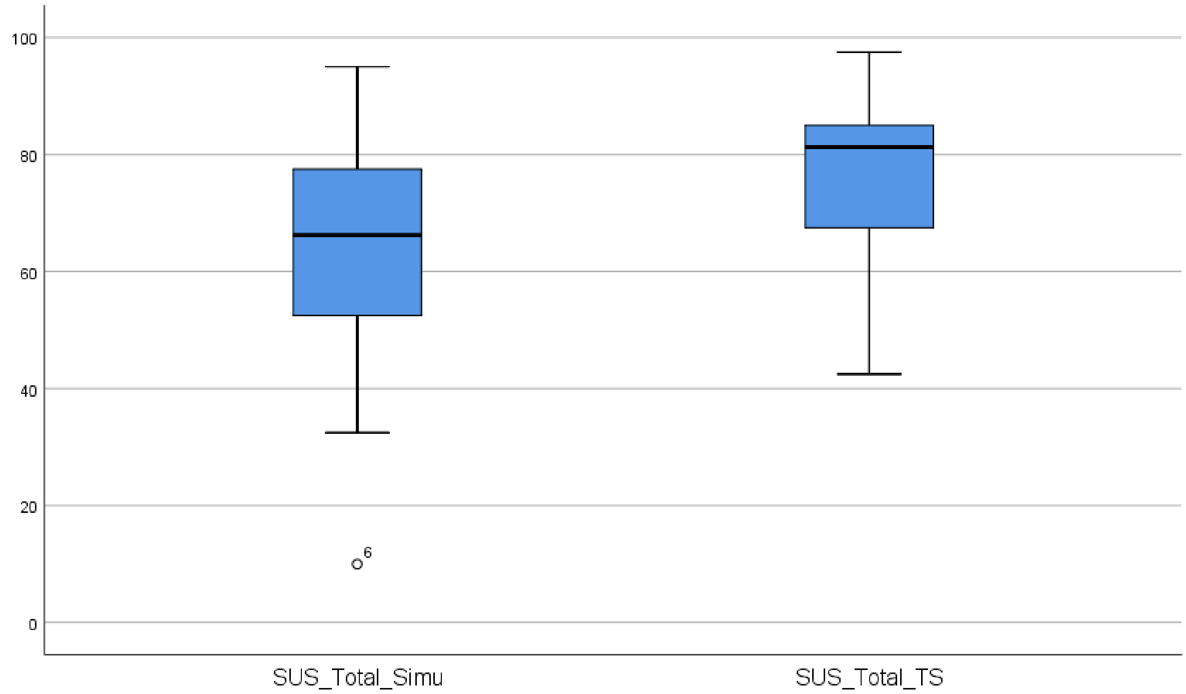

Figure 3: Boxplot of the SUS scores of the whole sample

The outlier respondent indicated in the comment box that she had difficulties with the SimuSurg application. However, she did not specify what kind of difficulties she experienced. Therefore, we cannot know if she had technical difficulties or if she just found that the app was difficult to use. In this study each analysis will be run in two datasets – one with 30 participants and one with 29 participants without the outlier. By default, the reported results will be those with the outlier included, but if an analysis of the second datasets has considerably different results, they will be mentioned.

Reliability check

In order to check the reliability of the SUS as a tool in the context of our research, a reliability analysis was done. Cronbach's alpha was chosen as a reliability estimate. Cronbach's alpha is a measure of internal consistency and it is one of the methods most widely used for estimating reliability. It provides a conservative estimate of reliability, and therefore received some criticism (Lewis, 2018).

There are well-established guidelines for acceptable alpha values for standardized questions, with an acceptable range from .7 to .95. Our analysis shows that the overall alpha of the SUS in the context of our research is .868. An optimum value of .9 is almost reached. This is in line with finding in literature which suggest that the SUS usually reaches coefficient alpha values between .7 and .9 (Lewis & Sauro, 2009). Therefore, the SUS can be accepted as a reliable tool to measure perceived usability in our research.

Assumptions for a linear regression models

To see if gaming frequency has predictive power for the SUS scores, the aim was to build two linear regression model with mobile gaming frequency as independent variable (IV) and SUS scores of TouchSurgery and SimuSurg as dependent variables (DV). Furthermore, two regression models with gender as IV and SUS scores as DVs was build. The assumptions that have to be fulfilled for linear regression analyses were checked and can be seen in Appendix E. The analyses of the assumptions showed that the distribution of the DVs is mostly normal. The analysis of the histogram of the residuals and the p-p plot of the residuals showed no unusual distribution of the residuals. Therefore, parametric statistics can be considered as suitable for the statistical analyses.

2.7 Data analysis

All data analyses will be performed with IBM SPSS statistics 26. The mean scores of the valid responses for TouchSurgery and SimuSurg will be calculated and compared. Afterwards, a linear regression model will be built to see if mobile gaming frequency has predictive power for the SUS scores. Since the mobile gaming frequency group consisted of three categories, a one-way ANOVA was performed to see if there are statistically significant differences in the means of the three groups.

To see if gender has predictive power for SUS scores of either of the two applications, a linear regression model was built. To see if the means of the gender groups are significantly different, an unpaired t-test was performed. It was chosen as the gender group consisted of two independent categories.

3.Results

3.1 Overall SUS ratings of TouchSurgery and SimuSurg

SUS scores for the 30 participants are depicted in Figure 4. TouchSurgery received better SUS scores overall (*M* = 77.00, *SD* = 13.69) than SimuSurg (M = 63.48 SD = 19.20*)*. Hence, the participants rated TouchSurgery better with 11.74 points on the SUS on average. According to the grading scheme proposed by Lewis and Sauro (2018), TouchSurgery receives a B as a grade and SimuSurg receives a C.

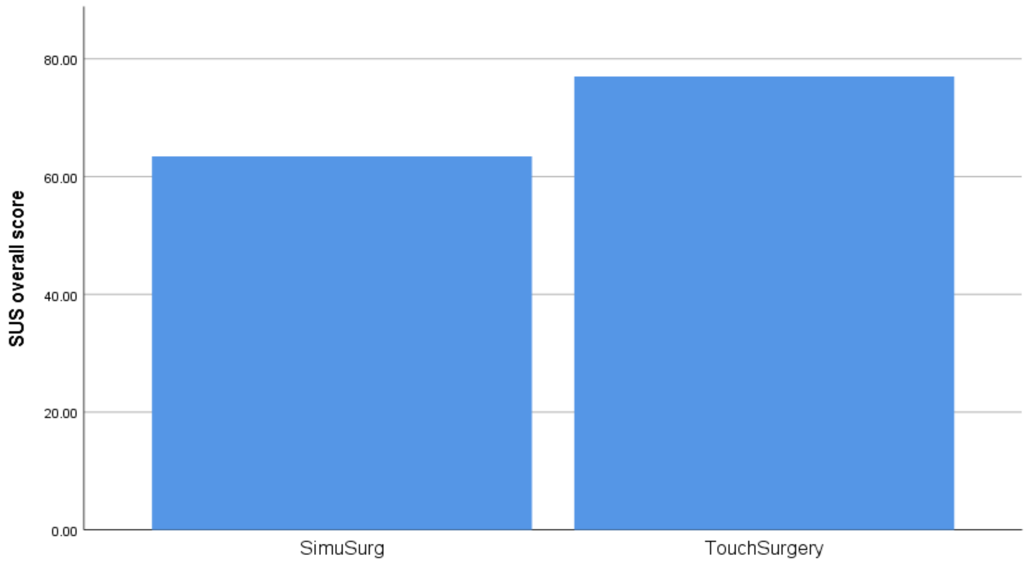

Figure 4: Mean SUS score of the whole sample

3.2 SUS scores grouped by mobile gaming frequency

The mean SUS scores of the two applications separated by mobile gaming frequency showed similar results among the three groups (figure 5). Participants who *never* play video games rated TouchSurgery the lowest among the groups (*M* = 74.32, *SD* = 15.21). In the *intermediate* group, participants rated the SUS score of TouchSurgery the highest (*M* = 80.91, *SD* = 10.45). The group playing video games *often* gave a rating comparable to the average of the whole sample for TouchSurgery (*M* = 77.14, *SD* = 17.22). The *never* group gave an average rating to SimuSurg (*M* = 65.00, *SD* = 17.42). The *intermediate* group rated SimuSurg best among all groups (*M* = 66.82, *SD* = *22.50*). The group playing video games *often* rated SimuSurg the lowest (*M* = 63.21, *SD* = 18.41). The results show that the mean SUS scores separated by gaming frequency do not show big differences among the groups.

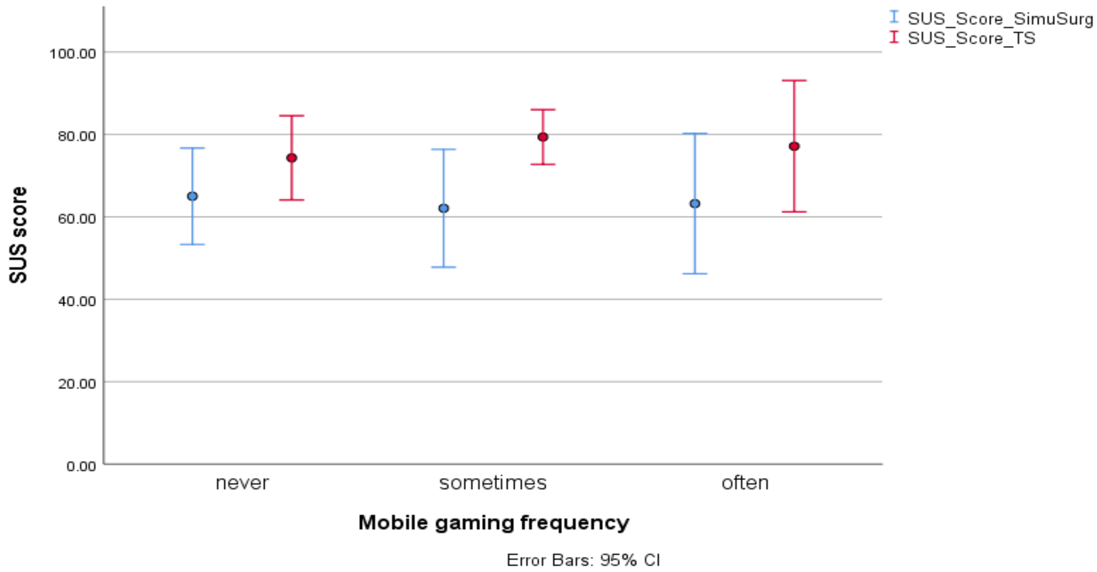

Figure 5: Error bars of the 95% confidence intervals. Dots representing means.

3.2.1 Mobile gaming frequency as predictor for user satisfaction

The results of the linear regression analyses for mobile gaming frequency as IV and SUS scores of TouchSurgery and SimuSurg as DVs are depicted in Table 3. Gaming frequency was not a significant predictor for SUS TouchSurgery scores ($F(1,28) = .277$, $p = .603$). For SUS SimuSurg scores gaming frequency also was not a significant predictor, either ($F(1,28) = .053$, $p = .819$).

Table 3.

Linear regression analysis with mobile gaming frequency as predictor

SUS Score	Variable	B	β	t	Sig.	95% Cl
TouchSurgery	(Constant)	73.74		11.016	.000	[60.03,87.45]
	Mobile gaming	1.746	.099	.526	.603	[-5.05, 8.55]
SimuSurg	(Constant)	65.43		6.945	.000	[46.13,84.73]
	Mobile gaming	-1.708	-0.44	-.231	.819	[-10.65,8.49]

Note: SUS_Score_TS R^2 = .01, SUS_Score_SimuSurg R^2 = .002

3.2.2 Comparing gamers vs non-gamers

To compare if the differences of the mean SUS scores separated by mobile gaming frequency are statistically significant, a one-way ANOVA was utilized. For the SUS TouchSurgery score, there was no significant difference between mobile gaming frequency groups ($F(2, 27) = .375$, $p = .691$). For the SUS SimuSurg score, no significant difference in mobile frequency groups was found, either ($F(2, 27) = .062$, $p = .691$).

Table 4.

One-way ANOVA of the mobile gaming frequency groups.

App		Sum of squares	df	Mean square	F	Sig.
SimuSurg	Between Groups	49.196	2	24.598	.062	.940
	Within Groups	10644.345	27	394.235		
	Total	10693.542	29			
TouchSurgery	Between Groups	146.944	2	73.472	.375	.691
	Within Groups	5295.556	27	196.132		
	Total	5442.500	29			

3.3 SUS scores grouped by gender

A comparison of the mean SUS ratings separated by gender is depicted in figure 6. Results show that the male participants rated SimuSurg (M = 69.69, *grade C*) approximately 13 points higher on the SUS than the female participants (M = 56.25, *grade D*). The male group also rated TouchSurgery (M = 79.84, *grade B*) better than the female group (M = 73.75, *grade B*) on the SUS with approximately 6 points.

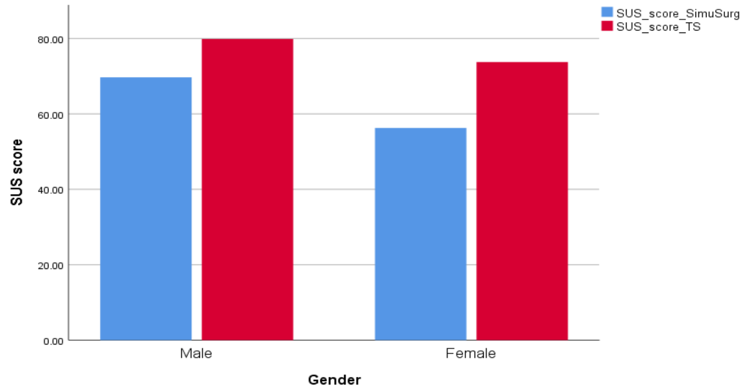

Figure 6: Mean SUS scores grouped by gender

3.3.1 Gender as predictor for user satisfaction

A simple linear regression was calculated to check if participant's SUS TouchSurgery score can be predicted by their gender. No significant regression equation was found (F(1, 28) = 1.503, p = .23). Another simple linear regression was performed to see if participant's SUS SimuSurg scores can be predicted by gender. Once again, the results were not significant (F(1, 28) = 4.04, p = .054).

Table 5.
Linear regression analysis with gender as predictor

App	Variable	B	β	t	Sig.	95% Cl
SUS_Total_TS	(Constant)	85.94		11.16	.000	[70.16, 101.71]
	Gender	-6.094	-.226	-1.23	.230	[-16.28, 4.09]
SUS_Total_SimuSurg	(Constant)	83.125		8.025	.000	[61.91, 104.34]
	Gender	-13.437	-.355	-2.010	.054	[-27.13,0.26]

Note: SUS_Total_TS R=.099, R^2 = .01, SUS_Total_SimuSurg R = .044, R^2 = .002

3.3.2 Comparing gender groups

To find out if the differences in the means of the gender groups are significant, an unpaired t-test was conducted. For the SUS TouchSurgery scores, the Levene's test for equal variance indicated that there is a significant difference in the variance (p = .045). The output of the T-test for non-assumed equality of variances indicates that there is no significant difference in

SUS TouchSurgery scores based on gender groups (t(23.37) = 1.20, p = .242).

For the SUS SimuSurg scores, equal variance was assumed in the Levene's test (p = .521). No significant effect of gender on the SimuSurg scores has been found (t(28) = 2.01, p = .054).

Table 6.

T-test to investigate significant differences between means of gender groups

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interv of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
SUS_score SimuSurg	Equal variances assumed	.422	.521	2.010	28	.054	13.43750	6.68582	-.25778	27.132
	Equal variances not assumed			2.012	27.594	.054	13.43750	6.67792	-.25067	27.125
SUS_score TouchSurgery	Equal variances assumed	4.404	.045	1.226	28	.230	6.09375	4.97053	-4.08792	16.275
	Equal variances not assumed			1.200	23.374	.242	6.09375	5.07936	-4.40442	16.591

4. Discussion

The aim of this study was to assess two applications related to learning MIS in terms of its perceived usability by different groups of people. To measure usability, the system usability scale was utilized. On average, TouchSurgery received better ratings in the SUS than SimuSurg. Neither mobile gaming frequency nor gender could be identified as predictors for the SUS scores of both applications. Although the differences in the ratings of the gender groups were noteworthy, the t-test showed that the differences were statistically not significant.

4.1 Gaming habits do not predict perceived usability

The claim of Isbister and Schaffer (2008) that frequent gamers perceive difficult games as more usable than non-gamers could not be observed in our context of educational games. Isbister and Schaffer (2008) highlighted that the difference in perceived usability between hardcore gamers and casual gamers depends on the difficulty of the game. In this study, no test was performed to check if participants perceive either of the two applications as a difficult game. It was assumed that SimuSurg might be perceived as more difficult due to its nature without further investigation. The Bedwell (2012) taxonomy showed that SimuSurg can be labelled as serious game, while TouchSurgery is a gamification with few game-like elements. Isbister and Schaffer (2008) wrote about the usability of games, but TouchSurgery may not even be perceived as a game by the user. To check the hypothesis of Isbister and Schaffer (2008), it may have been more accurate to compare two applications that both resemble a typical game but differ in difficulty and complexity.

Moreover, the nature of perceived usability and gaming frequency could be completely unrelated. A frequent gamer eventually performs better and learns faster in harder games (Isbister & Schaffer, 2008) as he/she is familiar with the concepts typically used in games, but the usability of a game may be assessed regardless of the performance. A person that performs well in a game could still believe that it is highly unusable.

Research about gaming habits and the influence on perceived usability is scarce in the literature. A study investigated the relationship of factors influencing technology acceptance of serious games (Wittland, Brauner, & Ziefle, 2015). Technology acceptance seems to be related to perceived usability to a certain degree (Holden & Rada, 2011; Scholtz, Mahmud, & Ramayah, 2016). In the study about the acceptance of serious games for healthcare and ambient assisted living environments, researchers found that it is independent of gender, technical expertise and gaming habits (Wittland et al., 2015). The authors also found that the gaming habits influenced the performance of the user in the serious games. The findings of the study are in line with the findings of our study that gaming habits do not influence the perceived usability of serious games. However, gaming habits seem to have an effect on the performance of the user in the serious game (Wittland et al., 2015).

4.2 Gender does not predict perceived usability

The mean SUS scores of TouchSurgery and SimuSurg grouped by gender show that the group of male participants in our sample perceived both apps as more usable than the female participants in our sample. Male participants gave higher ratings especially to SimuSurg. The regression analysis and the independent t-test were non-significant, regardless. Results on the effect of gender on the perceived usability are ambiguous in the literature. In a study about the influence of design aesthetics on perceived usability, no gender effects were found (Sonderegger & Sauer, 2010). In a study by Wittland, Brauner and Ziefle (2015), gender did not have any effect on the acceptance of serious games, either.

However, another study reported that men consistently rated the interface of car navigation systems better than the female participants (Lin & Chen, 2013). The authors suggested that the found effect could be related to a better sense of direction and a higher spatial confidence of males. This again may lead to a better perceived usability of car navigation systems (Lin & Chen, 2013). Although the t-test in our study was insignificant, we observed a trend that men consistently rated the usability of the applications higher. This is similar to the results found by Lin and Chen (2013). Both games and car systems are domains that are typically favored by male users (Isbister & Schaffer, 2008; Lin & Chen, 2013). Whether gender differences in perceived usability exist could depend on the type of system and how familiar the user feels with its use.

4.3 Usability assessment of TouchSurgery and SimuSurg by the whole sample

One of the aims of this paper was to assess the overall perceived usability of TouchSurgery and SimuSurg. TouchSurgery was considered more usable than SimuSurg. Since the approach of this study was of quantitative nature, it is hard to determine which features made the difference in perceived usability on an individual level. One way to find out about the reasoning of the participants for their ratings would be to look at the optional comments some of them left.

The better overall ratings of TouchSurgery also reflected in the comments. TouchSurgery received less comments overall, and most of the comments were related to the educational content of the application rather than its usability. Four participants reported that the application is too easy and may be useful for beginners and students. These comments were written by participants who were more advanced in the medical field. Only two participants commented on difficulties with the usability of the app. One person reported that he did not like the swiping, as it “does not add anything to the app”. Another one reported to have difficulties with the swiping, as well. Apart from that, no comments indicated difficulties with a satisfactory use of the application. For SimuSurg, more comments were written in general. Four people indicated that they had difficulties with the usability of the game, i.e. the camera movement, the proceeding to the next level or the utilization of the graspers. Two participants encountered bugs within the applications which forced them to restart certain levels. These kinds of comments indicate that some participants were rather dissatisfied with some aspects of the app, as they were not able to use it properly for various reasons. This may be due to the more challenging nature of SimuSurg in general. The game is intended as a sequence of challenging tasks, and not all users may enjoy the difficulty and the setbacks related to the game (Isbister & Schaffer, 2008).

The overall SUS scores and the feedback in the comment section indicate that TouchSurgery was designed in a more usable way. Participants encountered less technical difficulties. This is aggravated by the fact that an application like TouchSurgery does not allow as much freedom for errors compared to SimuSurg. TouchSurgery guides the user through the procedures and the user can hardly encounter difficulties throughout the process. SimuSurg utilizes a more challenging concept that can lead to complications and eventually to frustration which could lead to less perceived usability.

4.4 Strengths and Limitations

Usability testing of mobile applications can either be done in a laboratory setting or in a field study (Zhang & Adipat, 2005; Ma et al., 2011). Due to the limitations caused by the Covid-19 lockdown, the study had to be performed remotely as a field study. A field study has the advantage that the perceived usability of a mobile application is derived based on participants’ experience in a real environment, which is potentially more ecologically valid compared to laboratory experiments. Testing in a laboratory setting ensures that instructions are followed precisely, no irrelevant variables influence the participants and responses can be recorded to capture participants’ reactions. (Zhang & Adipat, 2005).

Although we tried to provide instructions that were as detailed as possible, we could not ensure that the participants fulfilled the tasks under similar conditions. They may have been distracted by their surroundings during the task, one participant fully focusing on the task while another one performed it while being distracted. Especially because the study was time-consuming, it is possible that some participants took a break while others did not. This can be assumed due to the time needed to complete the study, as some participants only took a little longer than 20 minutes while others took more than 80 minutes. The devices on which the participants performed the tasks were not standardized, as every participant had to do the task on their own phone. Mobile phone characteristics like screen size are proven to have an influence on the perceived usability of applications (Raptis, Tselios, Kjeldskov, & Skov, 2013). Having a singular phone model on which all the participants perform the tasks would make results more accurate.

Our sample was recruited via convenience sampling to ensure a rapid data collection, so the sample may not be representative for a larger population. Furthermore, we had a remarkable dropout rate, as only 30 out of 78 participants who opened the link fully completed the study. The detailed instructions and the long task durations seemed to be suboptimal for a remote study. We may have had dropouts who got overly frustrated with one of the applications and quit the study without indicating further. This would bias the usability of the app in a positive way, as it was not rated although it was perceived with a low usability. These kinds of dropouts can easily be tracked and integrated into the data in a laboratory setting, but for us it was not possible.

Finally, the categorization of the gaming groups was problematic in this study. Since this research was initially designed to assess perceived usability of different groups of medical expertise, the sample was not selected based on their gaming habits. Consequently, the category of gaming frequency was more of a by-product of the survey. Participants were asked how often they played mobile games per month/week, but there was no information about the duration of each gaming session. Some participants could have extensive gaming sessions once a week in which they play several hours while other participants may play three times a week for few minutes only. Nevertheless, the participant who plays for a few minutes three times a week would be considered a more frequent gamer than the person who plays several hours in one session in this paper. In Bedwell et al. (2012) for example, participants labelled “gamers” played an average of 17 hours per week. Categorizing people as gamers by the exact amount of time they devote to playing games would lead to more appropriate groups.

4.5 Future Research

It should be highlighted that the effect of gaming frequency did not show a significant effect on perceived usability in this specific study. In a more thoroughly planned setup that solely focusses on the gaming frequency as a research subject, different results may or may not be found.

As mobile applications as educational tools are relatively new, more research needs to be done in general. Next to the usability of the systems, we suggest investigating how useful the applications can prove as tools for learning compared to conventional tools. For more popular applications like TouchSurgery, studies have already confirmed its usefulness (Sugand et al., 2015). Similar studies about less popular applications like SimuSurg could prove helpful to distinguish between more and less useful applications. Furthermore, it is questionable if skills learned in applications like SimuSurg can be transferred to high-fi simulators or even to the real world to some degree.

4.6 Recommendations for educational practice

Mobile applications for medical educational are a concept that has potential (Franko & Tirrell, 2012). Medical experts that participated in our study suggested that TouchSurgery can be a useful tool especially for beginners, although it may be too easy for advanced healthcare workers. Since the overall SUS ratings of TouchSurgery were good, it can easily be recommended to people interested in the basics of medical education. TouchSurgery could be a great way to gain some additional theoretical input and to learn individually next to the training in the official curriculum. SimuSurg had more versatile feedback. Some participants liked the application and some participants disliked it. Two medical experts wrote that the application was not realistic and did not resemble real procedures, but it is questionable if the goal of the developers was to create a realistic application or if they just wanted to create a game through which players can get familiar with the very basic mechanics of MIS. Despite of some negative feedback for the applications, they are free of charge and available for everyone who possesses a smartphone. The biggest advantage of the applications is that people who are interested can try out the application without disadvantages or effort and then decide for themselves whether they perceive it as a useful learning tool.

4.7. Conclusion

The current work presented a usability test of two free mobile applications, TouchSurgery and SimuSurg. The assessment of the overall usability scores provided insights about the perceived usability of both applications. TouchSurgery was rated as a tool with a good score and SimuSurg as a tool with a mediocre score. Groups divided by mobile gaming frequency did not rate the applications significantly different and could not be determined as predictors for perceived usability of either of the applications. However, this may have been due to the premature setup of the research. Groups divided by gender were not significantly different either. Regardless, the difference in the mean values suggests that there may be a trend that men perceive the applications as more usable. As the gaming frequency and the gender do not seem to play a significant role for the perceived usability, the applications can be recommended to everybody and not just to a specific target group. Medical education on mobile applications can be viewed as a comfortable, cost-efficient addition for every user that seeks for uncomplicated ways to practice next to the traditional training methods.

Acknowledgements

I would like to thank my supervisors, Dr. Marleen Groenier and Dr. Simone Borsci, to make this project possible despite of the events of uncertainty and the setbacks that we experienced throughout the semester. Dr. Groenier's continuous support through regular meetings despite of the lockdown helped me to have thought-provoking impulses and to stay focused on the research. Her extensive feedback and her expertise in minimally invasive surgery were essential for the finalization of this thesis. Dr. Borsci's advice on the data analysis were valuable to find the right steps, and his feedback helped to improve this paper. I would like to thank them both for being patient with me.

Furthermore, I would like to thank my colleagues and friends Alexandru Amariei and Melina Kowalski who rapidly designed the setup of the research together with me in a short period of time. The cooperation with you was pleasant as per usual.

Finally, I would like to thank my family and my friends who always supported me throughout my years of studying. I am grateful to have you.

References

Aggarwal, R., Mytton, O. T., Derbrew, M., Hananel, D., Heydenburg, M., Issenberg, B., . . . Soper, N. (2010). Training and simulation for patient safety. *BMJ Quality & Safety, 19*(Suppl 2), i34-i43.

Alaker, M., Wynn, G. R., & Arulampalam, T. (2016). Virtual reality training in laparoscopic surgery: a systematic review & meta-analysis. *International Journal of Surgery, 29*, 85-94.

Arhippainen, L., & Tähti, M. (2003). *Empirical evaluation of user experience in two adaptive mobile application prototypes.* Paper presented at the MUM 2003. Proceedings of the 2nd International Conference on Mobile and Ubiquitous Multimedia.

Arnab, S., Lim, T., Carvalho, M. B., Bellotti, F., De Freitas, S., Louchart, S., . . . De Gloria, A. (2015). Mapping learning and game mechanics for serious games analysis. *British Journal of Educational Technology, 46*(2), 391-411.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, *24*(6), 574-594.

Basdogan, C., Sedef, M., Harders, M., & Wesarg, S. (2007). VR-based simulators for training in minimally invasive surgery. *IEEE Computer Graphics and Applications, 27*(2), 54-66

Beck, J. C., & Wade, M. (2006). *The kids are alright: How the gamer generation is changing the workplace*. Harvard Business Press.

Bedwell, W. L., Pavlas, D., Heyne, K., Lazzara, E. H., & Salas, E. (2012). Toward a taxonomy linking game attributes to learning: An empirical study. *Simulation & gaming, 43*(6), 729-760.

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing user satisfaction in the era of user experience: Comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. *International journal of human-computer interaction, 31*(8), 484-495.

Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), Usability Evaluation in Industry (pp. 189–194). London, UK: Taylor & Francis.

Brooke, J. (2013). SUS: A retrospective. Journal of Usability Studies, 8(2), 29–40.

Buttussi, F., Pellis, T., Vidani, A. C., Pausler, D., Carchietti, E., & Chittaro, L. (2013). Evaluation of a 3D serious game for advanced life support retraining. *International journal of medical informatics, 82*(9), 798-809.

Coursaris, C. K., & Kim, D. J. (2011). A meta-analytical review of empirical mobile usability studies. *Journal of usability studies, 6*(3), 117-171.

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011). Gamification. using game-design elements in non-gaming contexts. In *CHI'11 extended abstracts on human factors in computing systems* (pp. 2425-2428).

Dicheva, D., Dichev, C., Agre, G., & Angelova, G. (2015). Gamification in education: A systematic mapping study. *Journal of Educational Technology & Society, 18*(3).

Dieckmann, P., Friis, S. M., Lippert, A., & Østergaard, D. (2012). Goals, success factors, and barriers for simulation-based learning: A qualitative interview study in health care. *Simulation & gaming, 43*(5), 627-647.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*: sage

Franko, O. I., & Tirrell, T. F. (2012). Smartphone app use among medical providers in ACGME training programs. *Journal of medical systems, 36*(5), 3135-3139.

Frecker, M. I., Haluck, R. S., Dziedzic, R. P., & Schadler, J. R. (2007). *U.S. Patent No. 7,208,005.* Washington, DC: U.S. Patent and Trademark Office.

Gentry, S. V., Gauthier, A., Ehrstrom, B. L. E., Wortley, D., Lilienthal, A., Car, L. T., . . . Campbell, J. (2019). Serious gaming and gamification education in health professions: systematic review. *Journal of medical Internet research, 21*(3), e12994.

Gallagher, A. G., & Smith, C. D. (2003, September). From the operating room of the present to the operating room of the future. Human-factors lessons learned from the minimally invasive surgery revolution. *Seminars in laparoscopic surgery, 10* (3), 127-139.

Graafland, M., Schraagen, J. M. C., & Schijven, M. P. (2012). Systematic review of validity of serious games for medical education and surgical skills training. *The British journal of surgery, 99*(10), 1322-1330.
Groenier, M., Schraagen, J. M. C., Miedema, H. A., & Broeders, I. (2014). The role of cognitive abilities in laparoscopic simulator training. *Advances in health sciences education, 19*(2), 203-217.
Gurusamy, K., Aggarwal, R., Palanivelu, L., & Davidson, B. (2008). Systematic review of randomized controlled trials on the effectiveness of virtual reality training for laparoscopic surgery. *British Journal of Surgery, 95*(9), 1088-1097.
Hamilton, D., Lane, J. V., Gaston, P., Patton, J., Macdonald, D., Simpson, A., & Howie, C. (2014). Assessing treatment outcomes using a single question: the net promoter score. *The bone & joint journal, 96*(5), 622-628.
Harrison, R., Flood, D., & Duce, D. (2013). Usability of mobile applications: literature review and rationale for a new usability model. *Journal of Interaction Science, 1*(1), 1.
Holden, H., & Rada, R. (2011). Understanding the influence of perceived usability and technology self-efficacy on teachers' technology acceptance. *Journal of Research on Technology in Education, 43*(4), 343-367.
Huizenga, J., Admiraal, W., Akkerman, S., & Dam, G. t. (2009). Mobile game-based learning in secondary education: engagement, motivation and learning in a mobile city game. *Journal of Computer Assisted Learning, 25*(4), 332-344.
Hussain, A., Mutalib, N. A., & Zaino, A. (2014). *A usability testing on JFakih Learning Games for hearing impairment children.* Paper presented at the The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M).
Isbister, K., & Schaffer, N. (2008). The strange case of the casual gamer. In *Game usability: Advancing the player experience* (pp. 143-169). CRC press.
Jaffray, B. (2005). Minimally invasive surgery. *Archives of disease in childhood, 90*(5), 537-542.
Jordan, J.-A., Gallagher, A. G., McGuigan, J., McGlade, K., & McClure, N. (2000). A comparison between randomly alternating imaging, normal laparoscopic imaging, and virtual reality training in laparoscopic psychomotor skill acquisition. *The American journal of surgery, 180*(3), 208-211.
Kerfoot, B. P., & Kissane, N. (2014). The use of gamification to boost residents' engagement in simulation training. *JAMA surgery, 149*(11), 1208-1209.
Kortum, P., & Sorber, M. (2015). Measuring the usability of mobile applications for phones and tablets. *International journal of human-computer interaction, 31*(8), 518-529.
Landers, R. N. (2014). Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & gaming, 45*(6), 752-768.
Lewis, J. R., & Sauro, J. (2009). The factor structure of the System Usability Scale. In Kurosu, M. (Ed.), Human Centered Design, HCII 2009 (pp. 94–103). Heidelberg, Germany: Springer-Verlag.
Lewis, J. R. (2018). The system usability scale: past, present, and future. International Journal of Human–Computer Interaction, 34(7), 577-590.
Lewis, J. R., & Sauro, J. (2018). Item benchmarks for the system usability scale. *Journal of Usability Studies, 13*(3).
Lin, P.-C., & Chen, S.-I. (2013). The effects of gender differences on the usability of automotive on-board navigation systems–A comparison of 2D and 3D display. *Transportation research part F: traffic psychology and behaviour, 19*, 40-51.
Long, K. H., Bannon, M. P., Zietlow, S. P., Helgeson, E. R., Harmsen, W. S., Smith, C. D., . . . Group, L. A. I. (2001). A prospective randomized comparison of laparoscopic appendectomy with open appendectomy: clinical and economic analyses. *Surgery, 129*(4), 390-400.
Moreno-Ger, P., Torrente, J., Hsieh, Y. G., & Lester, W. T. (2012). Usability testing for serious games: Making informed design decisions with user data. *Advances in Human-Computer Interaction, 2012*.

Moss, G., Gunn, R., & Heller, J. (2006). Some men like it black, some women like it pink: consumer implications of differences in male and female website design. *Journal of Consumer behaviour, 5*(4), 328-341.
Novitsky, Y. W., Litwin, D. E. M., & Callery, M. (2004). The net immunologic advantage of laparoscopic surgery. *Surgical Endoscopy and Other Interventional Techniques, 18*(10), 1411-1419.
Olsen, T., Procci, K., & Bowers, C. (2011). *Serious games usability testing: How to ensure proper usability, playability, and effectiveness.* Paper presented at the International Conference of Design, User Experience, and Usability.
Raptis, D., Tselios, N., Kjeldskov, J., & Skov, M. B. (2013). *Does size matter? investigating the impact of mobile phone screen size on users' perceived usability, effectiveness and efficiency.* Paper presented at the Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services.
Reddy, P. K. (1994). Modular tools for laparoscopic surgery. In: Google Patents.
Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior, 69*, 371-380.
Scholtz, B., Mahmud, I., & Ramayah, T. (2016). Does usability matter? An analysis of the impact of usability on technology acceptance in ERP settings. *Interdisciplinary Journal of Information, Knowledge, and Management, 11*(2016), 309-330.
Sonderegger, A., & Sauer, J. (2010). The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. *Applied ergonomics, 41*(3), 403-410.
Soper, N. J., Brunt, L. M., & Kerbl, K. (1994). Laparoscopic general surgery. *New England Journal of Medicine, 330*(6), 409-419.
Spruit, E. N., & Band, G. P. (2014). Jaap F. Hamming & K. Richard Ridderinkhof. Psychological Research, 78, 878-891.
Stefanidis, D., Korndorffer, J., Markley, S., Sierra, R., & Scott, D. (2006). Proficiency Maintenance: Impact of Ongoing Simulator Training on Laparoscopic Skill Retention. *Journal of the American College of Surgeons, 202,* 599-603. doi:10.1016/j.jamcollsurg.2005.12.018
Surgicalscience (n.d.). LapSim essence sales guide. Retrieved from https://surgicalscience.com/wp-content/uploads/2017/09/Lapsim-Essence-Sales-guide.pdf
Sugand, K., Mawkin, M., & Gupte, C. (2015). Validating Touch Surgery™: a cognitive task simulation and rehearsal app for intramedullary femoral nailing. *Injury, 46*(11), 2212-2216.
Wang, R., DeMaria Jr, S., Goldberg, A., & Katz, D. (2016). A systematic review of serious games in training health care professionals. *Simulation in Healthcare, 11*(1), 41-51.
Wittland, J., Brauner, P., & Ziefle, M. (2015). *Serious games for cognitive training in ambient assisted living environments–a technology acceptance perspective.* Paper presented at the IFIP Conference on Human-Computer Interaction.
Zhang, D., & Adipat, B. (2005). Challenges, methodologies, and issues in the usability testing of mobile applications. *International journal of human-computer interaction, 18*(3), 293-308.

5. Appendix

Appendix A. Letter of invitation

Dear [],

We are three students from the psychology program of the University of Twente, Melina, Christof and Alexandru, and we are currently doing our bachelor's theses. The aim of our project is to test the **usability of mobile applications (serious games) used for training surgical skills**. We will look at how you perceive two apps and how you evaluate them. The goal of the study is to see how easy to use two applications are for different target groups. To achieve this, we want to receive some input from you as an end-user.

For us, the data which you will provide will be used in the writing process of our bachelor's theses and to inform the educational program about the usefulness of these kind of apps, for example for Endoscopic Skills. The benefit for you is to experience two applications through which you can train your surgical skills and learn about surgical procedures.

To complete the study, please make sure that you have a mobile device (smartphone) and a desktop computer or laptop available. You will have to test the applications on your mobile device and fill out a survey on the PC/laptop. The study will take approximately 45 minutes.

If you have any questions about participating in the study, do not hesitate to send us an email!

Click on this link to participate in the study:

https://utwentebs.eu.qualtrics.com/jfe/form/SV_39K1J0TeCFmUydT

Kindest regards,

The research team

Melina Kowalski, Christof Schulz, Alexandru Amariei

Appendix B. Information sheet

This usability-test represents a part of the project "Usability assessment of mobile applications used for training surgical skills". Your contribution will be used to evaluate two apps which are aiming to teach basic surgical skills. The goal of the study is to see how easy to use those apps are for different target groups. To achieve this, we want to receive some input from you, as an end-user. In this usability-test we will look at how you perceive the two apps and how you evaluate them. For us, the data which you will provide will be used in the writing process of our Bachelor's Theses. The benefit for you is experiencing two apps through which you can train surgical skills and learn surgical procedures.

During this session, you will have to perform tasks and answer questions:
• Firstly, we will ask for background information;
• Secondly, the actual usability-test will start. You will have to complete tasks in both apps. After each task, you will have to answer questions and upload proof of completion;
• Thirdly, you will receive questions about the session.

Below you can find some information about your rights and about the way in which your information will be handled:
• This session will take approximately 45 minutes. There is a limit of 30 minutes to successfully complete a phase, after which you can abort it and mention it in the questionnaire.
• You are free to withdraw yourself from this study at any given time, without providing a reason.
• For validation purposes, we will ask you to make screenshots to prove that you completed the tasks and upload them in the received form. Those screenshots should not contain any information that could be used to identify yourself.
• Your answers will be anonymized, safely stored, and accessed just by the members of the research team. If you decide at a later date that you do not agree with your data being used in the study, you can contact one of the researchers and ask for your answers to be removed without providing a reason.
• The applications you are going to test might use your personal data (e.g. device information).
• The Touch Surgery application will require you to create an account.
• The Touch Surgery application uses realistic depictions of medical procedures. Those depictions might be disturbing. If you do not feel comfortable with those depictions, you are advised to stop using the app and inform one of the researchers.

If you need further information about the research, before, during, or after the session, you can contact one of the researchers:
● Alexandru-Lucian Amariei (e-mail: a.amariei@student.utwente.nl);
● Melina Marie Kowalski (e-mail: m.m.kowalski@student.utwente.nl);
● Christof Schulz (e-mail: c.schulz-2@student.utwente.nl).

If you have questions about your rights as a research participant or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Secretary of the Ethics Committee of the Faculty of Behavioural, Management and Social Sciences at the University of Twente by ethicscommittee-bms@utwente.nl.

Appendix C. Consent form statements

1. I have read and understood the study information dated 03/06/2020, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.

2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.

3. I understand that taking part in the study involves:

- Providing some basic information about myself to the researchers' team;
- Testing two applications for training surgical skills;
- Completing and answering to the best of my ability to the questionnaires I will receive during the session;
- The applications I am going to use might also make use of some of the information I provide (e.g. results of the simulation).

4. I understand that information I provide will be used as input for evaluating two medical training applications and subsequently writing reports (Bachelor's Theses) about them.

5. I understand that personal information collected about me that can identify me, such as my age, gender or profession, will be anonymized and not be shared beyond the study team.

6. I give permission for the answers in the questionnaires that I provide to be archived in University of Twente student theses repository, so it can be used for future research and learning.

Appendix D. The System Usability Scale by Brooke (1996)

	Strongly disagree				Strongly agree
1. I think that I would like to use this application frequently.	○	○	○	○	○
2. I found the application unnecessarily complex.	○	○	○	○	○
3. I thought the application was easy to use.	○	○	○	○	○
4. I think that I would need the support of a technical person to be able to use this application.	○	○	○	○	○
5. I found the various functions in this application were well integrated.	○	○	○	○	○
6. I thought there was too much inconsistency in this application.	○	○	○	○	○
7. I would imagine that most people would learn to use this application very quickly.	○	○	○	○	○
8. I found the application very cumbersome to use.	○	○	○	○	○
9. I felt very confident using the application.	○	○	○	○	○
10. I needed to learn a lot of things before I could get going with this application.	○	○	○	○	○

As a minor change, *"system"* was replaced with *"application"*.

Appendix E. Task instructions

SimuSurg

This stage should take approximately 15 minutes. If you find yourself not able to successfully complete the task within 30 minutes, you can abort the task and mention it in the questionnaire.

Please read the instructions carefully and do not be afraid to take a second look in case you encounter a problem!

Task: Open the SimuSurg app. Press "**Start**". Now, press on "**Beginner**" and click on the first level named "**Scope introduction**". After looking at the instructions for the level, press "**Start**" once again. If you complete a level successfully, press "**Next activity**" and start the next level. Don't worry if you fail a level, you can simply re-try until you manage to solve it. **Please stop once you solved level no. 12, called "Irrigation Introduction" (in the "Intermediate stage")**. After completing the 12th level please take a screenshot.

Please do not forget to take a screenshot of the completion screen, after finishing the 12th level (Irrigation Introduction, in the Intermediate stage, seen in the bottom left corner of the screen). You can find instructions on how to do that below.

If you encounter a problem during this stage, please send an email to a.amariei@student.utwente.nl or a WhatsApp message at ....

After you completed the stage and answered the question at the bottom, you may proceed to the next section.

Touch Surgery

This stage should take approximately 15 minutes. If you find yourself not able to successfully complete the phase within 30 minutes, you can abort it and mention it in the questionnaire.

Please read the instructions carefully and do not be afraid to take a second look in case you encounter a problem!

Account set-up: To set up the account you will need to open the application and press on "**Create an account**". Fill in your email address and choose a password. Now you have to tick the first box to agree to the EULA, terms of agreement and privacy policy. The second box has to be ticked as well, to confirm that you are at least 18 years old. Now that you accepted the two necessary requirements, you can click on "**Create Account**" again. Press the "Find your procedures" to continue. You are now asked to fill in your first and last name and press "**Confirm**". You should see a page that asks for your profession. There are several options given to you, but you may also press "other/none of the above" at the bottom if none of them apply to you. Now, you will be asked what your main interests are. You can choose whatever you like or select one at random if none of them appeal to you. Your choice will not influence this research. After you chose your interests, you will be asked to indicate your secondary interest. Again, you can choose what you like or select one at random. You should be seeing the home screen of the application now.

<u>Task:</u> On the bottom of the page, you should see multiple icons. Please press the magnifying glass at the bottom of the page. If you press the correct icon you should be on a page with the search function on the top. Type in "**Laparoscopic Appendectomy**" in the search field. You should see a task with that name in the search results. If you press the task you should see a page with the option "**START LEARNING**". There are three learning and three testing phases. Please <u>only</u> complete the three learning phases. When you press "**START LEARNING**", the first learning phase should start. After finishing it you will see the options to exit, proceed with learning phase 2, or with testing phase 1. Please select "**Learn Phase 2**". After completing the second phase, you will have to repeat the same procedure to advance to the last stage, namely press on **"Learn Phase 3"**. After completing the third learning please, please take a screenshot.

Note: You <u>do not</u> have to complete the tests for each phase in this training course. We ask you to focus solely on the learning aspect of the course.

Please do not forget to take a screenshot of the completion screen, after finishing the 3rd learning phase (Appendectomy). You can find instructions on how to do that below.

If you encounter a problem during this stage, please send an email to <u>a.amariei@student.utwente.nl</u> or a WhatsApp message at ….

After you completed this stage and answered the question at the bottom, you may proceed to the next section.

Appendix F. Assumptions for linear regression

Assumptions for regression analyses with mobile gaming frequency as IV

For a linear regression analysis, the DV should be normally distributed. The distribution of the SUS scores of SimuSurg is normal (figure 7). The distribution of the SUS scores of TouchSurgery (figure 7) is a little skewed to the left but does not seem to violate the assumption of the normal distribution of the DV. The residual histogram and the p-p plot for SimuSurg indicate normality (figure 8). The residual histogram and the p-p plot of residuals look a little more scattered for TouchSurgery, but no pattern is visible that indicates an inappropriate fit of the parametric model (figure 9).

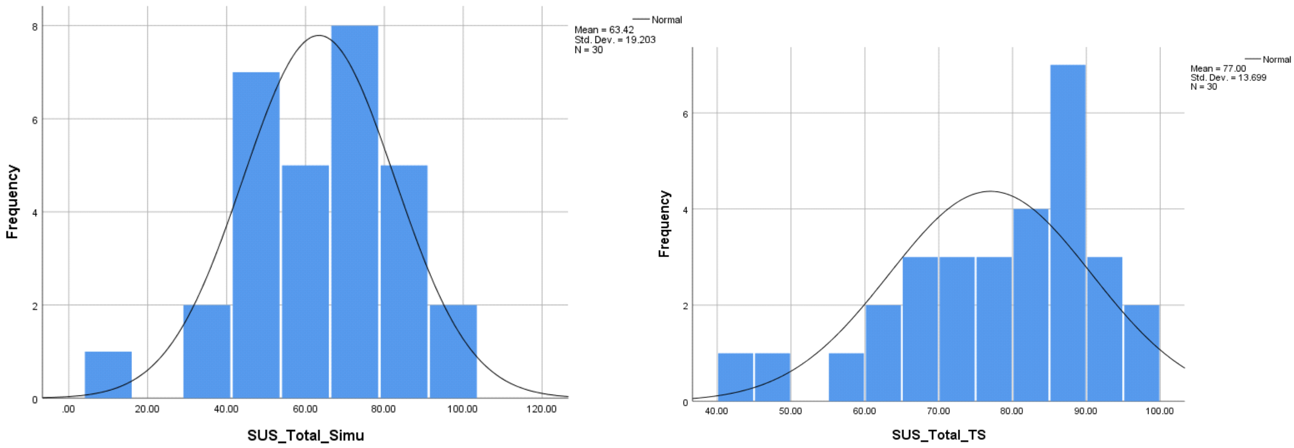

Figure 7: Distribution of SimuSurg (left) and TouchSurgery (right) scores

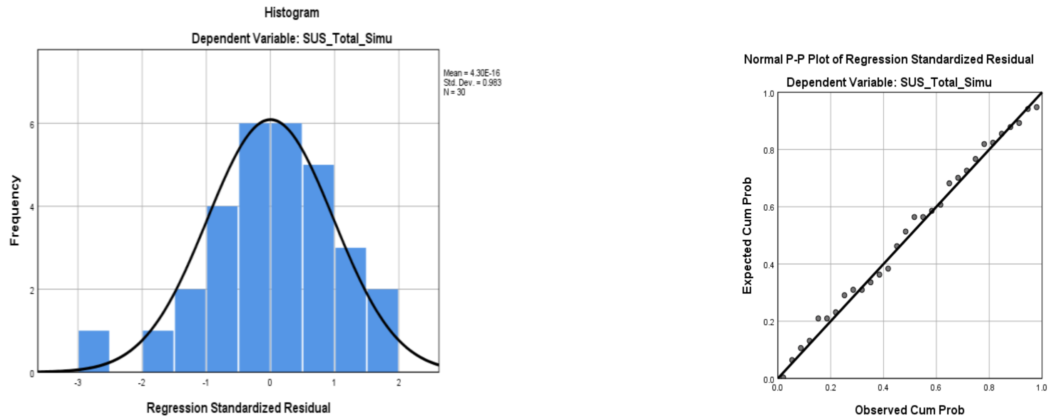

Figure 9: Residuals histogram (left) and p-p plot (right) for gaming frequency and SimuSurg

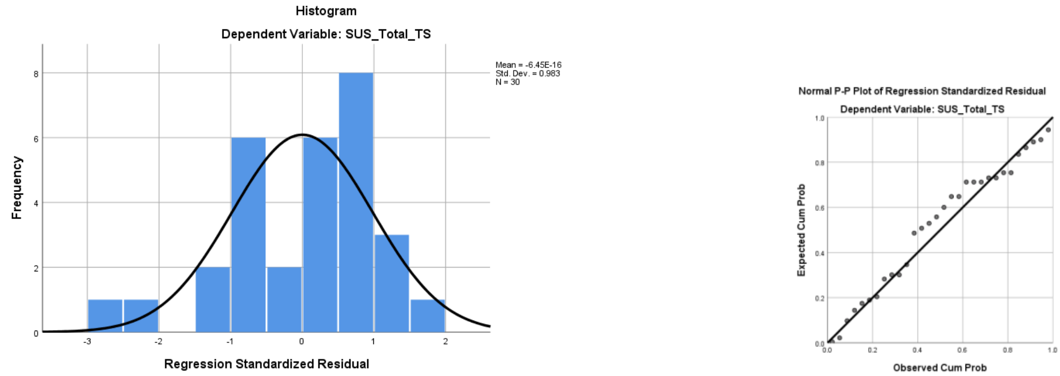

Figure 10: Residuals histogram (left) and p-p plot (right) for gaming frequency and TouchSurgery

Assumptions for the linear regression with gender as IV

As described above, the scores of TouchSurgery and SimuSurg are normally distributed (figure 7). The residual histogram for SimuSurg and TouchSurgery and the p-p plot for SimuSurg (figure 11) and TouchSurgery (figure 12) all indicate that the variables are multivariate normal and therefore suitable for linear regression.

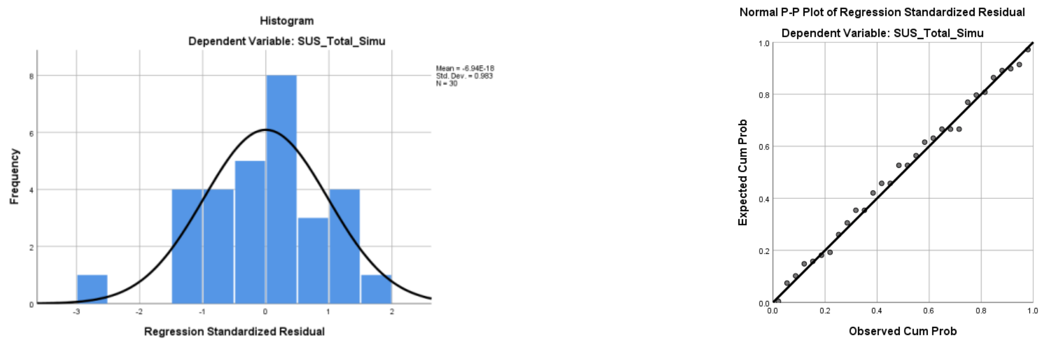

Figure 11: Residuals histogram (left) and p-p plot (right) for gender and SimuSurg

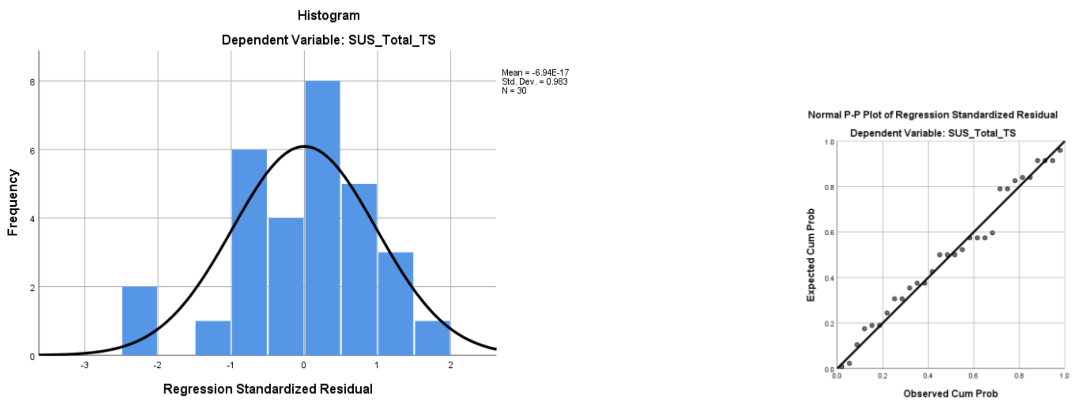

Figure 12: Residuals histogram (left) and p-p plot (right) for gender and SimuSurg