Usability assessment of medical training applications – Exploring the dimensionality of the System Usability Scale

Bachelor's Thesis
University of Twente
Department of Cognitive Psychology and Ergonomics
Faculty of Behavioral, Management and Social Sciences (BMS)
July, 2020

Student: Alexandru-Lucian Amariei
1st supervisor: Dr. Marleen Groenier
2nd supervisor: Dr. Simone Borsci

Foreword about the COVID-19 situation

Due to the situation created by the COVID-19 pandemic and organizational problems with the first thesis, this topic was proposed as a second-chance, short-timed alternative. This thesis was started on 02/06/2020 and finished on 24/07/2020. It was completed outside of the normal schedule of writing the Bachelor's Thesis at the University of Twente, with an extension granted by the Examination Board.

Nonetheless, the duration of completion was shorter (approximatively 8 weeks), compared to the regular semester normally allowed for writing the thesis. This short-time of completion and the restrictions created by the pandemic forced also a different planning of this study, with a shorter data collection period and a remote approach to conduct the usability tests.

Abstract

Background: Laparoscopic surgery represents an important practice in the medical field. There are several ways to train the skills required for it, for instance through the master-apprentice model, simulator training or serious games. The usage of serious games in medical education saw an increase in popularity, due to their value in learning. However, their perceived usability seems to be overlooked. It is important to consider the satisfaction of a system, as it influences the way in which it is perceived by the user. The System Usability Scale represents an instrument for measuring the perceived usability of a product. However, it was discovered that this scale might explore one more dimension: Learnability. The research on this subject is contradictory. Based on the study of Borsci, Federici, Bacci, Gnaldi, and Bartolucci (2015), the amount of exposure to a system seems to make this second component emerge. The aim of the present study is to see how another individual characteristic, namely the domain expertise, is influencing the structure of the scale. As systems to evaluate, from a perceived usability point of view, two applications which can be used in training laparoscopical skills were used: SimuSurg and Touch Surgery.

Methods: A within-subject design with the two applications presented in a random order was used. The study was conducted remotely, each participant completing it in their own time and place. The participants had to complete one task in each application and then rate their satisfaction using the System Usability Scale. Moreover, the participants were divided in three groups, based on their level of expertise: *novices*, *intermediates* and *experts*.

Results: The total scores of the System Usability Score showed a better perceived usability of Touch Surgery, compared to SimuSurg. The reliability analysis showed that the scale was reliable across all groups. Principal component analysis was performed on the data to see the effect of domain expertise on the dimensionality of the scale. An effect was visible in the *novice* and the *expert* groups, making the two-component structure emerge. Because of the inadequacy of the samples, the results of the individual applications and in the *intermediate* group were excluded.

Conclusions: Findings suggest that domain expertise has an effect on the way in which the subscales of the System Usability Scale behave. The two subscales were visible, however more ambiguously in the *expert* group, compared to the *novice* one. When looking at the satisfaction of the two applications, Touch Surgery seems to be more suitable for integration in the medical education. However, more research is needed to explore the link between the domain expertise and the structure of the scale and between user satisfaction and learning.

Table of Contents

1. Introduction

New surgical techniques, for instance, minimally invasive surgery (MIS), in combination with the shortcomings of the training curriculum of the beginner surgeons are creating a demand for new ways of training. Right now, training through simulators is an accepted way of familiarizing with the MIS procedures. However, serious games are gaining popularity as alternatives or adjutants in the training process. An aspect that should be explored when developing these new applications is their usability and the experience they create for the end-user. A good user-experience of the product stimulates the engagement of the users with the system. Therefore, to be adopted and successfully implemented in the curriculum of medical training, an application should also have an acceptable degree of usability. In this bachelor thesis, the usability of two applications used in medical training and the behaviour of the System Usability Scale, under the influence of the domain expertise, will be explored.

1.1. Training MIS skills

1.1.1. Skills necessary for MIS

Operating through MIS procedures requires surgeons to develop and apply different aptitudes, compared to the ones necessary for open surgery (Gallagher, Leonard, & Traynor, 2009). Those skills can be categorized into two major categories: psychomotor and cognitive. An example of a psychomotor skill necessary in MIS is using bimanual movements. This implies utilizing both hands to control two instruments at once (Hofstad et al., 2013). By doing this, the demand for using the non-dominant hand in synchronization with the dominant one is created, to be able to operate with both instruments concurrently. As a cognitive burden created by MIS, the two-dimensions (2D) to the three-dimensional (3D) mental conversion can be used as an example. This requires the practitioners to mentally convert the 2D representation they see on the screen to the real 3D representation of the operation space (Greco, Regehr, & Okrainec, 2010). There are several ways to train and develop these aptitudes.

1.1.2. Apprenticeship model for training

A way of learning the skills necessary for MIS is through the master-apprentice model. This approach involves letting the trainees observe and then practice MIS on patients, under the guidance of an experienced surgeon (Van Der Poel et al., 2016). This practice has two major flaws: it is highly dependent on the individual supervisor and on the amount of exposure to a variety of patients received by each trainee. Thus, consistent training and proficiency for all surgical trainees are not assured. Furthermore, a lack of available experienced personnel for training is existent in the medical field (Van Der Poel et al., 2016).

1.1.3. Simulator training

As a solution to the problems mentioned before, simulator training arises as an alternative to the traditional way of teaching. In the virtual reality (VR) simulator, a practitioner can experience the environment and train their psychomotor and cognitive abilities necessary for MIS, with minimal assistance from a third-party, for instance, a supervisor (Pierorazio & Allaf, 2009; Van Der Poel et al., 2016). Simulator-based training allows practicing in a standardized manner, objectively assessing the trainees' skills and reaching uniform proficiency levels after training (Dawe et al., 2014). Moreover, the acquired skills in the simulator are transferable to the real-world operating setting. Another advantage is the opportunity for the trainee to practice without risking human-lives, because of the artificial nature of the environment. People usually learn through making mistakes, approach which is not really an option in the apprenticeship model, but it is safely done in simulator training (Kneebone, 2010). Even if the high-fidelity simulators represent a valuable training tool, they also present a big barrier: economic accessibility. An MIS simulator can go into the range of thousands of euros to purchase. Those prices might not be accessible for every institution or practitioner. Therefore, other teaching methods were desired as precursors or alternatives for training MIS skills.

1.2. Serious gaming in medical training

A possible alternative for simulator training is using serious games (SG) for training. A SG is a digital application, including engaging gaming components, for instance, a ranking system, which is keeping the user interested and simultaneously offer them knowledge or skills relevant for the real world (Graafland, Schraagen, & Schijven, 2012). Simulators are great for training psychomotor skills and improving the dexterity of the trainees. However, through SGs, a broader set of skills could be trained, including the prerequisite knowledge about the medical procedures that have to be performed.

SGs are different from regular games. They include a learning component, which is differentiating them from conventional games. Moreover, they are also suitable for training procedures in the medical field, due to their ability to simulate and support complex decision-making processes. Graafland et al. (2012) mention that SGs are engaging the users in the learning process, because of the challenges they are creating. Thus, through completing the different stages of a game to reach the objective, the player is learning without experiencing too much effort. This balance between learning, challenge and the engagement they create for the players are making SG valuable training methods. SGs seen an increase in development and usage for teaching medicine, because of the value they provide (Graafland et al., 2012). However, even if the teaching component is explored, it does not seem that too much emphasize is put of their usability aspects.

1.3. Usability and user experience

The usability of a system is a measurement regularly employed in the human-computer interaction (HCI) area. It represents the degree to which a product can be used to reach a specific objective, by the specific user in a specific usage context (International Organization for Standardization, 2018). Furthermore, in the process of reaching the goal, the system should employ three characteristics, to be considered usable: effectiveness, efficiency, and satisfaction for the user. Meeting those criteria is conferring the product the ability to be easy to learn, captivating and efficient (Kaya, Ozturk, & Gumussoy, 2019).

Going a step further, user experience (UX) is a concept that includes usability but also goes beyond it (Petrie & Bevan, 2009). With the emergence of new systems and technologies, it is considered that users are not looking anymore just for the characteristics which are defining usability. They also seek pleasant aesthetics and being captivated by the product. Borsci et al. (2015) are elaborating more on that, mentioning that there is a general agreement that the experience of a user with the product is influenced by their perceived usability, the looks of the system and the degree to which their requirements are met. As it was mentioned earlier, user satisfaction represents a partial measurement of usability, and implicitly UX. It usually correlates with efficiency and effectiveness, but it can also be explored independently from them. Usually, measuring the perceived satisfaction is done through questionnaires, looking at how a user is subjectively evaluating the system.

1.4. The System Usability Scale

1.4.1. A short history of the System Usability Scale

A questionnaire used to evaluate the perceived satisfaction of a product is the System Usability Scale (SUS). During the years, it was applied in a multitude of tests and it is considered a *de facto* standard when measuring the user's satisfaction with a system (Lewis, 2018). This instrument was developed by Brooke (1996), as a "quick and dirty" usability measurement. It contains ten simple statements about the system's usability. Each statement can be evaluated on a five-point Likert scale (ranging from *Strongly disagree* to *Strongly agree*). The tone of the items alternates from positive to negative, odd-numbered questions being positive-toned and even-numbered questions negative-toned. The SUS score for a system can range from 0 to 100. However, this score alone does not hold any meaning.

As Lewis (2018) is mentioning, the meaning of a score emerges when there is comparison. For instance, when obtaining the SUS score of a system, it is necessary to use it as a benchmark and compare it to the score of another system. This has to be done to be able to draw conclusions on which system is more satisfactory for the user. As more SUS data was generated over the years, it was possible to elaborate some norms. These norms, applied through additional scales, were developed to better interpret the scores given by SUS, without the need for comparing two different systems. Notably, Bangor, Kortum, and Miller (2009) developed an adjectival scale, based on the scores obtained in SUS. This tool can be used as an attachment to get a better perspective on how the SUS score of a system will translate in words (ranging from *Worst Imaginable* to *Best Imaginable*). The adjectival rank is given based on the range in which the total SUS score belongs. Following normative research is the work of Sauro and Lewis (2016). In their studies, they found that the average of the SUS is 68. Based on this and on computing the percentile ranks for the whole range on SUS scores, the Curved Grading Scale (CGS) was created. This instrument employs grades (letters ranging from *F* to *A+*) assigned depending on the range in which the SUS score fits. Because of the eleven grade ranges of CGS, compared to the seven ranges of the adjectival scale, CGS is offering a more refined method to determine the usability of a system (Lewis, 2018). SUS is often used because of its high reliability and proven validity, somewhat short size, and low cost (Bangor, Kortum, & Miller, 2008). Those aspects are making it a popular option for practitioners in the field of HCI when investigating the perceived usability of a system.

1.4.2. The dimensionality of the System Usability Scale

SUS was initially created to be unidimensional, evaluating just the perceived usability of a system. However, in 2009 it was found through factor analyses of SUS answers that Items 4 and 10 are loading onto a different component, compared to the rest of the items (Borsci, Federici, & Lauriola, 2009; Lewis & Sauro, 2009). Lewis and Sauro (2009) termed the subscales Learnability (Items 4 and 10) and Usability (the rest of the items). However, the following studies from 2009 to 2015 were not successful in replicating this two-factor structure

(Lewis, 2018). The SUS seemed to not be robust enough to elicit the two components under all circumstances.

In the study of Borsci et al. (2015), the two subscales of SUS were again observed. In their study, the level of experience of the user with the system was considered a condition that made this bi-dimensionality visible. They determined that the second dimension appears if the user has sufficient exposition with the product. However, the attempt to replicate this study failed (Lewis, 2018). During a more recent study on a large sample of SUS questionnaires, Lewis and Sauro (2017) concluded that the bi-dimensionality of the scale is aligning with the tone of the items and not with the previous two identified subscales. Therefore, it was recommended to treat the scale as unidimensional when extracting its meaning, instead of considering two-components. Nonetheless, Borsci et al. (2015) suggested that bi-dimensionality could emerge because of time of exposition and the expertise gained with a product. SUS behaved as a unidimensional scale when administered to individuals who had less product experience and became bi-dimensional when used with users with more experience. The present study will explore the way in which the SUS acts under another characteristic: the domain expertise.

1.5. Research goal

In this paper, the aim is to see how the individual expertise in a domain, is influencing the dimensionality of the SUS. For this, two applications were selected (SimuSurg and Touch Surgery) in the domain of medicine. The applications' usability will be evaluated by three groups of people using the SUS. Each group has a different level of expertise in the medical domain: *novices*, *intermediates*, and *experts*.

1.6. SimuSurg and Touch Surgery

One application which could be used in training is SimuSurg. SimuSurg is a SG for mobile devices (e.g. smartphones, tablets) simulating basic psychomotor tasks in a MIS environment (Royal Australasian College of Surgeons, n.d.). It was developed in collaboration with the Royal Australasian College of Surgeons (RACS) and it targets potential surgeons. The game is structured on four levels of difficulty, each level encompassing six tasks. Those simulations are covering essential aspects from MIS, for instance, employing the endoscopic tools to grasp objects. The participant is receiving video and written instructions before performing each task and also a rating afterward, based on their performance. The tasks require the users to use the phone gyroscope to look around the environment and the touchscreen to maneuver the tools. At a superficial look, it can be said that the application is covering some of the skills necessary in MIS. Namely, some of the tasks are requiring participants to concurrently use both thumbs to move the endoscopic tools and mentally convert the 2D environment they see into 3D. However, there are no studies done to see if the application is a valid training tool for prospective surgeons.

Another application used for medical training is Touch Surgery (TS). This application is aimed at helping medical personnel to learn different procedures and test their knowledge (Digital Surgery Limited, n.d.). It offers a wide range of medical procedures from various medical specialties, each course being developed by a medical institution. The courses are usually divided into parts, starting with the equipment preparation and concluding with the actual procedure. Moreover, each part is divided into the simulation and the test. The user has to follow the simulation first, which includes graphical representations, interactive or not, of the procedure's steps and written instructions on what they are supposed to do. After finishing the simulation, the user can complete a test, to evaluate their understanding of the procedure they studied. The nature of the simulations and the interactivity are offering TS gamification characteristics, to keep the users engaged. Compared to SimuSurg, TS's validity was already tested (Kowalewski et al., 2017). Its usefulness and validity were confirmed by users and experts and is considered a good adjutant in the training process for medical trainees.

2. Methods

2.1. Study design

This study used a within-subject design, with two applications presented in a random order. It was an online-based usability-test and it was conducted remotely. Participants received the study instructions and had to complete it by themselves, in their own time at a place of their own choosing. They were able at any time to contact a member of the research team for further clarification. This approach was chosen because of the social and physical interaction limitations created by the COVID-19 pandemic in 2020. The study was approved by the Ethical Committee of Behavioural and Management Sciences (BMS) at the University of Twente (request no. 200884).

2.2. Participants

In total, 79 participants voluntarily participated in the study (Figure 1). After applying the inclusion criteria, 45 entries were excluded because of not completing the whole study, two because they used TS before, one because of completing the study in less than ten minutes and one because they encountered technical problems with SimuSurg and was not able to rate the application. In the end, 30 entries were considered valid. The participants were then divided into the three main groups. From the *other* category, two individuals were included in the *novice* group and two in the *intermediate* one. Following that, 17 individuals were assigned in the *novice* group (7 females [41.2%], M_{age} = 21.8, SD = 2.31), 6 in the *intermediate* (3 females [50%], M_{age} = 24, SD = 0.63) and 7 in the *expert* (4 females [57.1%], M_{age} = 36.14, SD = 7.4).

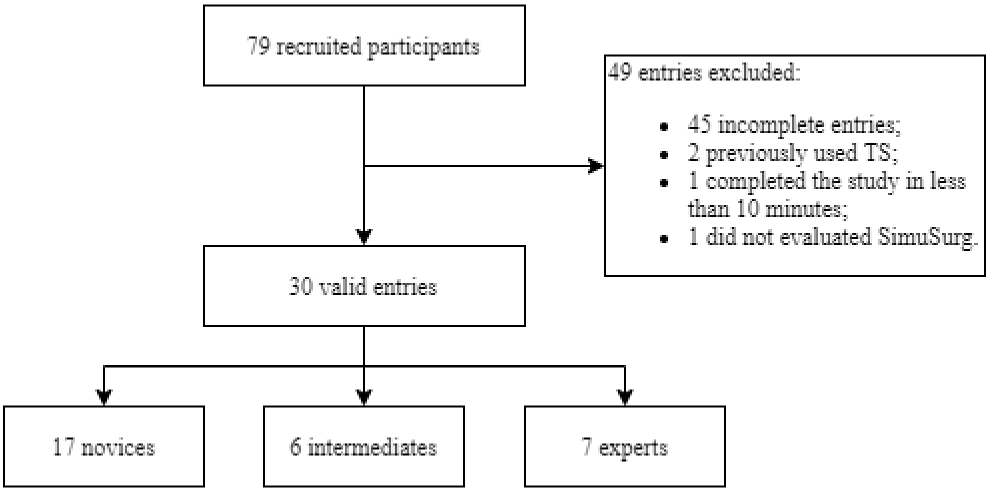


Figure 1. Participants in the study. The participants were checked based on the inclusion criteria and then divided into the three groups.

Convenience sampling was used as a method to recruit the participants. The participants were divided into three groups, based on their experience in the field of medicine. The first group was the *novices*, comprised of students without a medical background. The second group was the *intermediates*, including medicine students with a minimum of one year of study in the area. This one-year experience time was chosen to assure that the students are already accustomed to the basics of medical procedures. The last group is the *experts*, comprising surgeons or nurses with at least two years of experience in the field. This time was chosen because it is considered that they had the opportunity to already experience several medical procedures. Additionally, the order in which the tasks in the two applications were shown was randomized. The idea of the randomization there was to avoid allowing the participants to compare the second application with the first one.

Several inclusion criteria for all groups were also implemented. The participants had to possess a smartphone/tablet and a personal computer to be able to take part in the study. They were required to not have any previous experience with the two applications: SimuSurg and TS. Also, they were required to provide proof of completion of the tasks, either in the questionnaire through a file upload or through eyewitness testimony to one member of the research team. This measure was used to ensure that the participants completed the required tasks and the number of tasks necessary to assume that they can properly evaluate the usability of the applications. Lastly, they were required to completely fill in the survey in a reasonable amount of time. It was considered that at least 20 minutes are necessary to complete the tasks in both applications and the questionnaire. Also, the survey had a 24 hours' limit to submit the response. Therefore, any response outside this interval was invalidated.

2.3. Materials

2.3.1. Informed consent

The informed consent was composed of three parts (see Appendix A). The first one was a short invitation letter to which the link to the survey was attached. The second component was the information brochure, within the questionnaire, including the aim of the study, the benefits for both parties, the rights of the participants, and contact details of the researchers and the Ethics Committee of BMS, at the University of Twente. The final part was comprised of six statements that the participant had to agree or disagree with.

2.3.2. Demographics questionnaire

The questionnaire was administered through the Qualtrics platform and included questions about the participant's age, gender, mobile device model, and current occupation. Also, two questions about the frequency they play mobile and video games, with four options were included (ranging from *regularly* to *never*).

2.3.3. Instruction sections

Several instruction sections were included in the survey. Those elements were essential because of the remote nature of the study. Their purpose was to offer the participants instructions on how to successfully set-up the study and complete it, with minimum contact to another party (e.g. member of the research team). Before the tasks have begun, the participants were instructed on how to set-up (install) the applications on their devices. During the tasks, they had instructions on how to take a proof of completion (screenshot). Also, in the TS app, they received instructions on how to set-up an account. Another instruction section was placed in the file upload page, to give the participants instructions on how to upload the screenshots they took. Finally, at the end of the survey, they received instructions on how to uninstall both applications and delete the TS account. Graphical representations were also used in some instruction sections, to facilitate the process for the participants.

2.3.4. System usability scale and additional scales

The SUS was placed after the task in each application. The scale used was the original one, developed by Brooke (1996), with a minor change in the terminology of the statements. Namely, the word *system* in the 10 statements was replaced with *application*, to better fit the context (see Appendix B). Equation (1) depicts the method for computing the SUS score for each participant.

$$\text{SUS Score} = [(\text{SUS1} - 1) + (5 - \text{SUS2}) + (\text{SUS3} - 1) + (5 - \text{SUS4}) + (\text{SUS5} - 1) + (5 - \text{SUS6}) + (\text{SUS7} - 1) + (5 - \text{SUS8}) + (\text{SUS9} - 1) + (5 - \text{SUS10})] * 2.5 \tag{1}$$

Moreover, the GCS developed by Sauro and Lewis (2016) was used to assign grades to the scores of SUS. This was mainly done to see where the scores are placed, below average, average, or above average. The grades and the correspondent SUS score intervals, starting from the lowest to the highest, are as follows:

- Grade F (0–51.7)
- Grade D (51.8–62.6)
- Grade C– (62.7–64.9)
- Grade C (65.0–71.0)
- Grade C+ (71.1–72.5)
- Grade B– (72.6–74.0)
- Grade B (74.1–77.1)
- Grade B+ (77.2–78.8)
- Grade A– (78.9–80.7)
- Grade A (80.8–84.0)
- Grade A+ (84.1–100)

2.3.5. Net Promoter Score

The Net Promoter Score (NPS) was included as an additional measurement of the participants' satisfaction of the applications. NPS represents a standardized measure used in determining customer satisfaction with a product (Krol, de Boer, Delnoij, & Rademakers, 2015). It is comprised of a question: "How likely are you to recommend [the product/company] to a friend or a colleague?" and a Likert scale, ranging from 0 (*not at all likely*) to 10 (*extremely likely*). Looking at the scores, it can be determined how many detractors (ratings between 0 to 6), passives (ratings from 7 to 8), and promoters (ratings from 9 to 10) each application might have.

2.3.6. Post-task questionnaire

A post-task questionnaire was also included after each task. It was composed of three questions with dichotomous answer options. The participants were asked if they used the respective application before, if they successfully completed the task and if they encountered any problems that impeded their progress with the stage. Lastly, they had a text-entry box for adding any additional comments about the stage.

2.3.7. End-of-survey questionnaire

The end-of survey questionnaire included one multiple choice question and a text-entry box. The multiple-choice question asked the participants about which application they preferred more and offered four choices (*both, SimuSurg, Touch Surgery,* or *none*). The text-entry box was included to offer the possibility of adding remarks or comments about the session.

2.4. Procedure

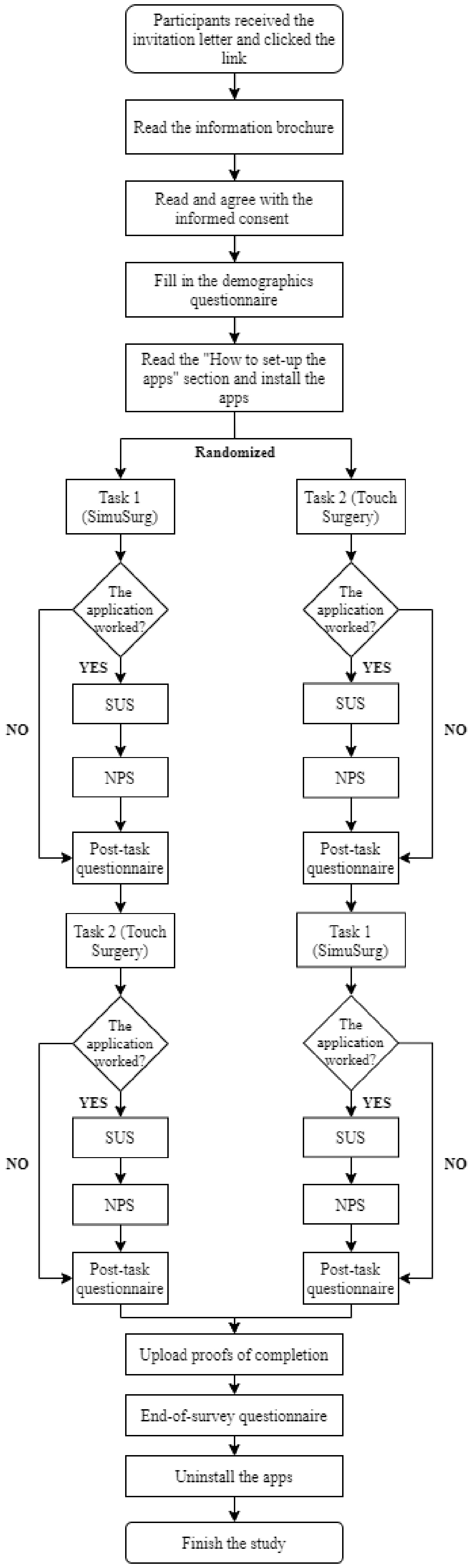

Participants received the invitation letter and clicked the link
Read the information brochure
Read and agree with the informed consent
Fill in the demographics questionnaire
Read the "How to set-up the apps" section and install the apps
Randomized
Task 1 (SimuSurg)
Task 2 (Touch Surgery)
The application worked?
YES
NO
SUS
NPS
Post-task questionnaire
Task 2 (Touch Surgery)
Task 1 (SimuSurg)
The application worked?
YES
NO
SUS
NPS
Post-task questionnaire
Upload proofs of completion
End-of-survey questionnaire
Uninstall the apps
Finish the study

Figure 2. Flowchart depicting the procedure of the study. To be followed from top to bottom.

The procedure for this study is depicted in Figure 2. All participants were invited to the study using the invitation letter, with the link to access the survey attached to it. When opening the link, the information brochure was displayed. After reading it, the participants had to proceed on the informed consent page, where they had to agree or disagree with the statements. The next step was completing the demographics questionnaire. When this was done, the participants encountered the *How to set-up the apps* page. After installing both apps, they were able to proceed to the first task. The two tasks were randomized. However, in this thesis SimuSurg will be presented as the first and TS as the second. The participants had to complete the task in SimuSurg and confirm that they were able to run the application on their device. If the application worked properly, they proceeded to complete the SUS and the NPS. If not, they skipped those two sections and were redirected to the post-task questionnaire. After completing the post-task questionnaire, the participants were able to proceed to the second task, in TS. As an additional step before the task, the participants had to set-up an account to be able to use TS. The following steps are the same as in the SimuSurg. When they were done with the post-task questionnaire for the second task, the participants were instructed on how to upload the proofs of completion of the two tasks and invited to do it. After uploading the screenshots, they were asked to fill in the end-of-survey questionnaire. Lastly, the participants received information on how to uninstall both applications and delete the account from TS. After this step, the study was done.

2.5. Tasks

Each participant had to perform two tasks during this study. One task was carried out in the SimuSurg application and the other one in the TS application. The paths that each participant had to follow to complete the tasks are depicted in Figure 3. Small variations for progressing are present and dependable on the participants' preference and phone model. However, the flowcharts illustrate the ideal, most common paths that can be followed. Each task will be described in more detail in the text below. Furthermore, the instructions for each task can be found in Appendix C.

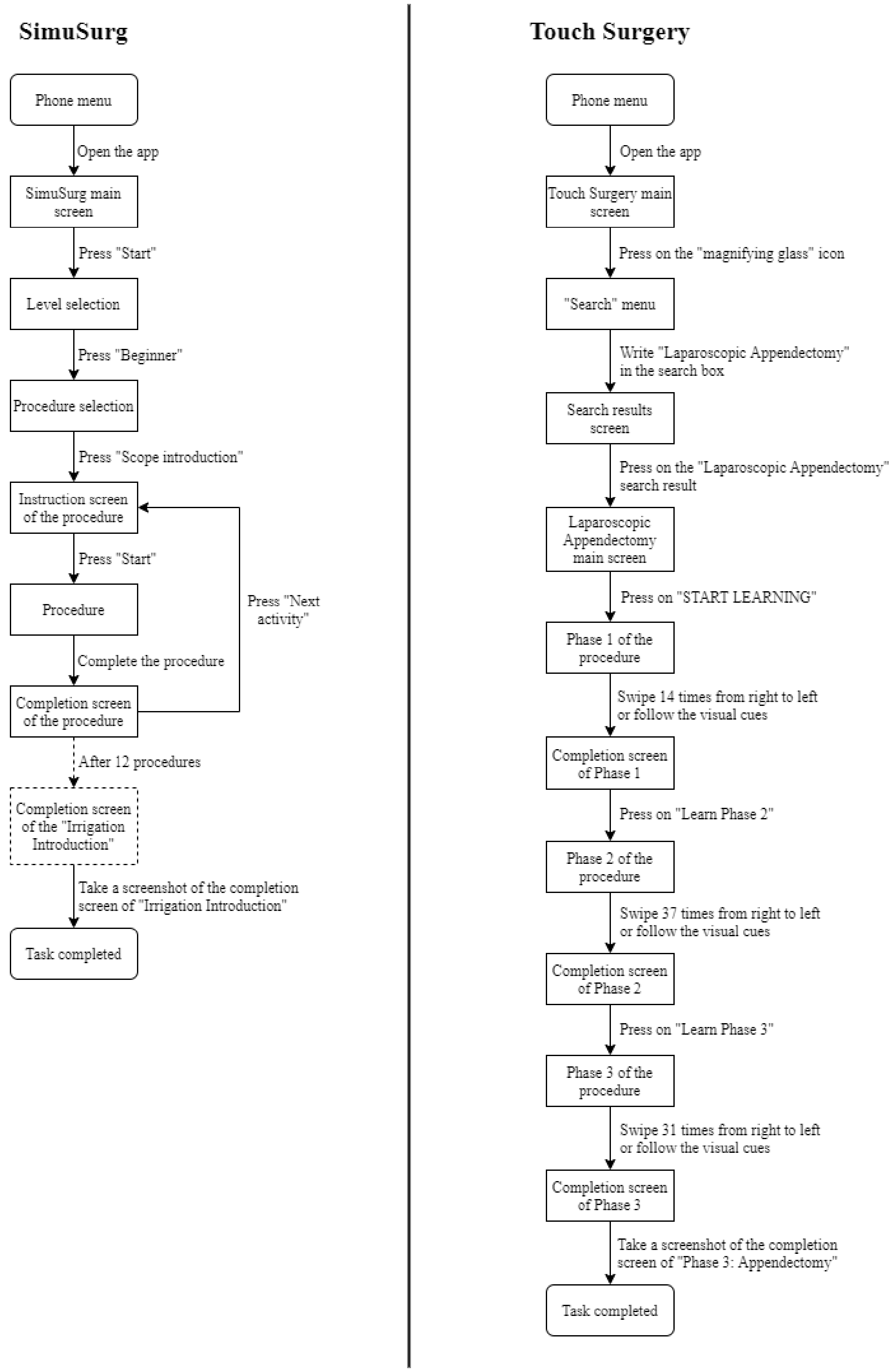

SimuSurg
Phone menu
Open the app
SimuSurg main screen
Press "Start"
Level selection
Press "Beginner"
Procedure selection
Press "Scope introduction"
Instruction screen of the procedure
Press "Start"
Procedure
Press "Next activity"
Complete the procedure
Completion screen of the procedure
After 12 procedures
Completion screen of the "Irrigation Introduction"
Take a screenshot of the completion screen of "Irrigation Introduction"
Task completed
Touch Surgery
Phone menu
Open the app
Touch Surgery main screen
Press on the "magnifying glass" icon
"Search" menu
Write "Laparoscopic Appendectomy" in the search box
Search results screen
Press on the "Laparoscopic Appendectomy" search result
Laparoscopic Appendectomy main screen
Press on "START LEARNING"
Phase 1 of the procedure
Swipe 14 times from right to left or follow the visual cues
Completion screen of Phase 1
Press on "Learn Phase 2"
Phase 2 of the procedure
Swipe 37 times from right to left or follow the visual cues
Completion screen of Phase 2
Press on "Learn Phase 3"
Phase 3 of the procedure
Swipe 31 times from right to left or follow the visual cues
Completion screen of Phase 3
Take a screenshot of the completion screen of "Phase 3: Appendectomy"
Task completed

Figure 3. Flowcharts depicting the expected progress within the two apps. To be followed from top to bottom. The diagram on the left shows the task progress in SimuSurg and the one on right the progress in TS.

2.5.1. SimuSurg task

The aim of the task in the SimuSurg was to let the participants experience the mini-games depicting elemental MIS procedures. For this reason, the first 12 procedures were chosen, this covering the *Beginner* and *Intermediate* levels offered by the application. It was assumed that completing those will give the participants from all groups an adequate experience in the SimuSurg.

After installing SimuSurg, the participants had to open it from the phone menu. Then, the main screen of the application appeared. By pressing *Start*, the *Level Selection* menu was opened. SimuSurg requires first users to complete a task to unlock the following one. Therefore, the participants had to press on *Beginner*, to be able to select the first procedure. After selecting it, through pressing on the *Scope Introduction*, the participants were redirected to the instruction screen. To start the procedure, they had to press on *Start*. Each mini-game had different requirements, which needed to be completed in a certain amount of time to be considered successful. When the participants successfully completed the procedure, a completion screen appeared. To proceed to the next mini-game, they had to press on the *Next activity* button. The participants had to repeat this string of actions until they reached and successfully completed the 12 th procedure: *Irrigation Introduction*. When this procedure was completed, the task in SimuSurg was considered successfully passed. As proof of completion for this task, the participants had to take a screenshot of the completion screen of the *Irrigation Introduction* screen. The screenshot showed the time of completion and the obtained score.

2.5.2. Touch Surgery task

The aim of the task in TS was to see how participants perceive the application and the learning experience it offers. TS offers a wide range of courses for medical procedures. For this study, the *Laparoscopic Appendectomy* was chosen. This decision was made because of the fit with the MIS area and the mild graphical representation of the procedure. It was decided that it is better to select a procedure with milder graphics, as a realistic one might have not been suitable for all groups. The task included solely the learning aspect of the procedure. The participants were asked to focus on the simulation of the procedure and the information they receive and skip the testing phases. This was done because of the time constraints and to avoid frustrating the participants. Before the actual task, the participants were instructed on how to set-up an account. However, this aspect was considered a prerequisite and is not included in the task description.

After installing TS and setting the personal account, the participants had to open the application and go to the main menu. From the main menu, they had to press on the *magnifying glass* icon at the bottom of the screen to access the *Search* menu. In the *Search* menu, they had to tap on the search box from the top and write in *Laparoscopic Appendectomy*. Once they wrote it, they received a results screen with the procedure on it. To access the course, they had to press on the search result. The course screen showed details about the course and the phases. To start from the beginning, they had to press the *START LEARNING* button. After doing this, they were redirected to the 1 st learning phase of the course. To finish, they had to swipe from the right to left 14 times or use the interactions within the simulation. After progressing through all steps, a completion screen of the phase appeared. To advance to the 2 nd phase, they had to press *Learn phase 2*. The process of progressing in the 2nd and 3rd phase was the same as in the first one, the only difference being the number of steps. After completing the 3 rd phase, the participants encountered the completion screen. At this point, the task was considered successfully passed. As proof of completion, the participants had to take a screenshot of the completion screen showing that *Phase 3: Appendectomy* was finished.

2.6. Data analysis

Descriptive statistics (mean, standard deviation) were used to summarize the demographics and to see how the SUS scores looked in all groups. The SUS scores were computed for all participants, in both applications. Following, grades from the CGS, developed by Sauro and Lewis (2016) were assigned for a better overview. To be able to see the reliability of the questionnaire, the even-numbered statements were converted from negative to positive. Furthermore, Cronbach's Alpha was computed for the SUS questionnaires in both applications and in the aggregated dataset, to examine its reliability.

Firstly, it was necessary to determine if the samples are adequate for factor analysis. To examine this, the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) and Bartlett's Test of Sphericity were computed for both applications in each group. After determining if the samples' sizes are sufficient, principal component analysis (PCA) was performed for the answers in the two applications, across the three groups. The purpose of this analysis was to see how the Items of the SUS are loading based on the level of expertise. This was done initially without imposing a factor loading, to see how many components are extracted based on the eigenvalues (higher than 1). Then, a forced two-factor structure was used to see if the factor loadings in the two components are matching the two subscales of the SUS. Afterwards, it was decided to perform a PCA by aggregating all the data and using both the SUS questionnaires filled by each participant. This approach was considered valid in this situation because of the main aim of the study, which implies looking at the scale and not particularly at the usability of the applications. By that, it is meant that we wanted to see how the individual differences across the groups are influencing the factorial structure of SUS, independently from the application. All analyses were performed using IBM SPSS Statistics 24.

3. Results

3.1. The perceived usability of the two applications

3.1.1. SUS scores in the three groups

The SUS ratings for both applications, independent and across the three groups, can be seen in Table 1. The mean values suggest a better perceived usability of TS, compared to SimuSurg along all groups. In the *novice* and *expert* groups, this difference was larger than in the *intermediate* group. This is supported also by the attributed grades, both applications having *C* as a grade in the *intermediate* group. Overall, TS's usability can be considered better than SimuSurg's. Looking at the standard deviation (SD) values, the scores varied the most when rating the SimuSurg application, in the *novice* and *expert* groups.

Table 1

The mean values of the two applications independent from groups and the means, standard deviations and grades of SUS in the two applications, across the three groups.

Group	Independent (n = 30)	Novices (n = 17)			Intermediates (n = 6)			Experts (n = 7)		
Application	Mean	Mean	SD	Grade	Mean	SD	Grade	Mean	SD	Grade
SimuSurg	63.41	60.73	21.3	D	67.5	14.57	C	66.42	18.64	C
TS	77.00	79.11	13.4	A-	70.41	13.36	C	77.5	14.93	B+

3.1.2. Reliability of the questionnaire

Before exploring the reliability of the SUS measurements, the answers undergone refactoring. Specifically, the even-numbered questions, which are negative-toned, were changed to a positive-tone. Table 2 shows the results of the reliability analysis with and without considering the domain expertise, in both applications and in the aggregated dataset. When performing the reliability analyses on the answers, without considering the group distributions, the results were acceptable. Cronbach's Alpha values shown values above 0.7 for each application, in each group. Furthermore, the results when employing the repeated-measures structure on the data were above 0.8. These values are in line with the acceptable norms for standardized questionnaires and with previous research on the SUS (Lewis, 2018).

Table 2

Reliability analyses on the two applications and the aggregated dataset

Group	Independent	Novices	Intermediates	Experts
Dataset	Cronbach's Alpha	Cronbach's Alpha	Cronbach's Alpha	Cronbach's Alpha
SimuSurg	.868	.901	.878	.887
Touch Surgery	.787	.748	.832	.856
Aggregated	.886	.877	.848	.875

3.2. The dimensionality of the SUS in the two applications

3.2.1. SimuSurg

KMO and Bartlett's Test were performed first to see if the samples in each group are adequate. To be considered acceptable, KMO for a sample should normally be above .500 and the Bartlett's Test should produce values lower than .05 to recommend factor analysis. In the *novice* group, the sample was considered marginally acceptable, with a KMO = .572 and a .000 value of Bartlett's Test. However, those analyses were not useful for the *intermediate* and *expert* group, as the correlation matrix was not positive definite. This might have been caused by the negative eigenvalues in the two groups. PCA was still used for the dataset, firstly with extracting components represented by eigenvalues higher than one. This resulted in 3 extracted components for the *novice* group, 2 for the *intermediate* group and 4 for the *expert* one. When imposing a two-component structure, the bi-dimensionality of the SUS was slightly visible in the *novice* and *expert* group, but not in the *intermediate* one. Varimax rotation was used during all analyses.

3.2.2. Touch Surgery

KMO and Bartlett's Test were performed, with unsatisfactory results for the TS data. KMO was .479 and Bartlett's Test was .055 for the *novice* group. For the other two groups, the output was the same as for the SimuSurg analysis. Because of this, the dataset for TS was considered inconclusive. Nonetheless, PCA was still performed and its results can be found in Appendix D.

3.3. The dimensionality of the SUS after restructuring the dataset

3.3.1. Restructuring the dataset

The analyses of the individual applications along the 3 groups were considered inconclusive. This was decided because of the inadequacy of the samples, based on the values of KMO and Bartlett's Test. Furthermore, some groups suggested a four-component structure, which is highly improbable when considering the theoretical underpinnings of the SUS. The scale was initially designed as unidimensional; therefore 4 components are too much to be considered valid. As a solution, the data was aggregated. A repeated-measures structure was chosen instead, using each participant's both SUS questionnaires' answers (from SimuSurg and TS).

3.3.2. PCA on the aggregated dataset

The same procedure as for the individual applications was used. The adequacy of the sample was explored across the three groups. KMO and Bartlett's Test showed improved values for the *novice* group (KMO = .766, Bartlett = .000) and for the *expert* one (KMO = .631, Bartlett = .000), compared to the individual application analyses. The *intermediate* group results were successfully computed with this dataset but it did not recommend factor analysis, as the sample was not adequate (KMO = .457, Bartlett = .063). Nonetheless, PCA was performed for all groups, but just the results for the *novice* and *expert* group are reported. The results for the *intermediate* group are included in Appendix D. Varimax rotation was used, in accordance with similar previous research (Borsci et al., 2015; Lewis & Sauro, 2017). When not forcing a number of factors, 2 components were shown for the *novice* group and 3 components for the *intermediate* and *expert* groups. The results after imposing the two-component structure with the Varimax rotation can be seen in Table 3.

Table 3

Results of the PCA in the novice and expert groups. Items are in their original order from the questionnaire and the loads for the two components are attributed to each group.

Group	Novices		Experts	
Items	Usability	Learnability	Usability	Learnability
1	.739		.645	
2	.700		.756	
3	.831		.456	.506
4		.769		.723
5	.811		.670	.592
6	.577		.652	.627
7	.523	.613		.795
8	.787		.777	
9	.733		.790	
10		.807		.922

Note. The loads which did not match the expectations or were not controversial were removed for clarity purposes. The complete table, including the *intermediate* group, can be found in Appendix D.

3.3.3. Novice group

When imposing the two-component structure on the *novice* group, the components of the SUS were discernable. The Usability component was the first, showing higher loads between the items ranging from 1 to 3 and from 5 to 9. The Learnability component emerged secondarily, as Items 4 and 10 had the highest load (.769 and .807 respectively). However, item 7 shown a splitting between the two factors, with similar loads in both of them (.523 in Usability and .613 in Learnability).

3.3.4. Expert group

After imposing the two-component structure, the bi-dimensionality of the SUS became slightly visible in the *expert* group. For the first component, the Usability, all items loaded, excepting 4, 7 and 10. In the Learnability component, Items 4 and 10 had a relatively large loading between them, compared to the other items. This might suggest that the bi-dimensional structure of SUS also appeared in the *expert* group. However, some problems were identified. Items 3, 5 and 6 loaded at a reasonably high level in both components, instead of loading just in the Usability subscale, according to expectations. Moreover, item 7 loaded unexpectedly high in the Learnability component and not in the Usability. Looking at the content of the 7 th statement of SUS (Appendix B), it can be assumed that also this item has a relation with the Learnability component.

4. Discussion

The main aim of this study was to see if domain expertise is influencing the emergence of the two-component structure of the SUS. Previously, it was shown that the amount of exposure to a system is making the two subscales visible (Borsci et al., 2015). A SG and an application employing gamification elements, aimed at training MIS skills, were used as systems that had to be evaluated by the participants. Three groups with different levels of expertise were selected, starting with the *novice* group, comprising inexperienced individuals, *intermediates* with knowledge in the medical domain and *experts* with the most knowledge in the field, compared to the other groups.

4.1. The dimensionality of the SUS

The outcome of the study suggests that there might be a link between the domain expertise and the bi-dimensionality of the SUS. In the study of Borsci et al. (2015), the subscales were visible when the participants had an enough amount of exposure to the system. Therefore, it was assumed that other factors might also make this two-component structure emerge, namely the domain expertise. Based on the results of the study, it seems that domain expertise is somewhat influencing how the SUS's structure behaves. It was observed that the two components are indeed visible for the *novice* group, even if also Item 7 loaded in both Usability and Learnability. However, the results in the *expert* group were marginally relevant in supporting the existence of the subscales. An effect was visible, based on the factor loads in the subscales. Even though Items 4 and 10 did not load in the first component and loaded reasonably in the second one, there were some anomalies. Namely, the manner in which Items 3, 5 and 6 loaded was contradictory to the expectations. Furthermore, Item 7 behaved in an unexpected manner, with a much higher load in the Learnability subscale.

It was not clear why Items 3, 5 and 6 loadings split between the two components in the *expert* group. However, for Item 7, in both groups, it can be assumed that the high load appeared because of its content ("I would imagine that most people would learn to use the application very quickly"). Moreover, SimuSurg's tasks are designed for learning psychomotor skills and for TS the participants were specifically instructed to focus on the learning phase of the procedure. Therefore, the evaluation of satisfaction with the two applications differed from evaluating a regular system, for instance a webpage or a software. The participants might have focused on the degree of which the applications facilitate learning, after one usage. These aspects might explain why Item 7 loaded unexpectedly high into the Learnability subscale, instead of Usability. In the study of Borsci et al. (2015), the bi-dimensionality of the SUS was visible after the participants had more exposition of use with the product. Compared to that study, the trend between the two groups in this study shows that the two-component structure emerged in the opposite direction. Individuals in the *novice* group had the lowest level of domain expertise and the ones in the *expert* group had the highest experience in the domain. However, the two subscales were clearly visible in the group without experience in the domain and appeared ambiguously in the most experienced group. These aspects represent interesting points of reflection and exploration for future studies.

The results for the individual applications and the *intermediate* group were declared inconclusive. Firstly, it was also planned to see how the individual differences in expertise are influencing the structure of SUS for the two applications. However, because of the unsatisfactory initial analyses, the results were removed. The main problem with these data seems to be the insufficient sample size. After aggregating the data in a repeated-measures structure, a viable dataset was produced. It is important to note an aspect of the new dataset. SimuSurg is a SG aiming at training psychomotor skills necessary for MIS, when TS is a training application, aimed at training cognitive skills, employing some gamification elements. Even if both are designed for training medical practitioners, they are aiming at training different skills. This difference might have created some cofounding in the new dataset, which might explain some of the differences in the items' loads. However, previous studies successfully used large aggregated SUS answers of different systems to explore the dimensionality of the scale (Lewis & Sauro, 2017). Therefore, it is hard to conclude if the difference of the learning objectives in the two applications really affected the results.

The findings of more recent studies on the dimensionality of SUS seems to partially hold. Brooke (1996) developed the scale as a unidimensional measurement of the perceived usability of a system. Afterwards, Lewis and Sauro (2009) and Borsci et al. (2009) observed that the scale might actually have a two-component structure, looking at Usability and Learnability, finding that was refuted by subsequent studies. Borsci et al. (2015) proposed that the amount of exposure to a product might influence the way in which SUS behaves. However, Lewis, Utesch, and Maher (2015) failed to replicate the study. Lewis and Sauro (2017) conducted a large study on the structure of SUS and concluded that the scale should be used as an unidimensional measurement, as it was originally intended. The results of our study were not clear enough neither to confirm nor to infirm this suggestion. Even if some effects were present, it was not sufficient to clearly determine how the level of expertise in the domain is influencing the two subscales of SUS. It is important to note that this study was explorative and inconclusive. An effect seems to exist, but it was not sufficiently visible to definitively conclude if the SUS is uni- or bi-dimensional, under certain circumstances. This means that more research should be done to see how the scale behaves under the influence of domain expertise, and maybe other factors.

4.2. The usability of SimuSurg and Touch Surgery

As a side aim, the usability of SimuSurg and TS was explored. Based on the scores and the grades from the CGS, elaborated by Sauro and Lewis (2016), it can be considered that TS was considered more usable by the *novice* and *expert* groups. SimuSurg received grade *D* in the *novice* group and *C* in the *expert* one, which suggests a barely level of usability. For TS, the grade was *A-* in the *novice* group and *B+* in the *expert* category. The *intermediate* group rated the applications similarly, with a corresponding *C* grade for both of them.

It is difficult to determine why those differences appeared. This is because the study was not specifically designed for exploring the perceived usability aspect of the individual applications. However, some insight can be obtained from looking at the participants' comments. For SimuSurg two participants form the *novice* group reported that they were confused by the camera movement created when moving the phone. Two other participants reported that task six was annoying to perform and one mentioned that the game was unnatural, not understanding its purpose. Also, one participant from the *novice* and one from the *intermediate* group encountered technical bugs in the application, which made them restart a level. An expert reported that the manipulation of instruments in the game was difficult and three others mentioned that they do not see its link to the real-life procedures. It can be assumed that the lower rating of SimuSurg in the first two groups was influenced by the application's technical problems, when for the *expert* group the value of the application in medical training was not recognized.

In TS, the comments focused more on the un-intuitive swiping animation of the application. Five participants across the three groups reported that they had problems when trying to progress through the procedure. In the *expert* group, two participants reported that they see the value of the application for trainees when two participants also reported that it was too easy for them. Even with some problems, from a user satisfaction standpoint, it seems like TS might be more useful in the training process than SimuSurg. As some of the experts mentioned, TS could be valuable as a tool to familiarize with the medical procedures. Moreover, even if it was not tested in this study, the application also offers the possibility to test the knowledge about the studied procedure. This might add more value in the medical educational context, compared to SimuSurg. Nevertheless, more research should be done to explore how the user satisfaction of these applications is influencing the learning process of the trainees.

4.3. Limitations

The first limitation of this study is the context in which it was performed. The COVID-19 pandemic and the short time in which the study was done forced a remote approach to collect data. This resulted in completely relying on participants correctly setting-up the study environment, following the instructions, and reporting the results. Some methods to verify the validity of the responses, for instance, the submission of a proof of task completion, were required. However, it was still not as reliable as observing the participants directly.

The second limitation was created by the exclusion of the *intermediate* group. This decision was taken based on the inadequacy of the sample, based on the initial analyses. It would have been interesting to also see how the domain expertise of a group with more knowledge than the novices but with less than the experts is influencing the structure of SUS. Unfortunately, the constrains of the study did not facilitate the recruitment of a large enough sample to be considered representative.

4.4. Recommendations

Even if this study was not conclusive with replicating the two-component structure of SUS, some effects were still visible. We focused solely on the domain expertise as an individual characteristic which may influence the dimensionality of the scale. Future studies might employ a longitudinal study design, exploring how the domain expertise and the longer or repeated exposure to a system are influencing the internal structure of SUS. Furthermore, it is recommended to gather large enough samples, to properly perform the factor analysis and see if the two subscales are present.

Looking at the two applications used in this study, we can say that TS has a better degree of perceived usability than SimuSurg. This does not mean that SimuSurg holds no value in medical training. However, based on the results, TS seems more suitable for being used in the training process of medical trainees. It would be interesting to see how this application is being perceived when included in the curriculum of medical students or starting medical practitioners. Another point of interest for the future is the relation between user satisfaction and learning. It would be interesting to explore how TS, and even SimuSurg, are facilitating the learning process and how is this related to the perceived usability they create for the user.

5. Conclusions

The SUS is still one of the most used and appreciated measurements for perceived usability in the industry. Even if it might not necessary also explore a second aspect of a system, the Learnability, it is a valuable and reliable tool. This study's findings are not discrediting the theory that SUS evaluates two components. However, the results are not conclusive enough to support this theory either. Domain expertise seems to have an effect on the structure of the SUS, but it is not clear to which degree. More research should be conducted to see how domain expertise is influencing the subscales of the SUS, before deciding if the scale should be used in its original unidimensional form or not. Looking at the applications used in this study, TS seems to have more value as a learning tool for medical practitioners. However, its relationship with learning and the relationship between perceived usability and learning represent something to explore in the future.

Acknowledgments

I would like to thank dr. Marleen Groenier, who provided the amount of help and initiative which made this thesis possible. Her knowledge in laparoscopically research and the constructive feedback offered helped me in getting a better understanding of the topic and coherently putting the information and my thoughts in this paper. Moreover, the continuous motivation offered during the short period in which this thesis was written was invaluable for its completion.

I would also like to thank dr. Simone Borsci, for sharing his knowledge regarding the usability domain. The study conducted by him and his colleagues in 2015 represented the cornerstone of this thesis. Moreover, the help with analyzing the data and interpreting the results was exceptionally useful in the writing and the reflection process.

Also, I would like to thank my colleagues, Christof Schulz and Melina Kowalski, with whom the design of this study was created and who provided their input and opinions about the study whenever it was necessary.

Lastly, but not least, I want to sincerely thank my parents for all the unconditional support offered thorough the whole process of writing this thesis and to the people close to me, who took some of their time to offer constructive criticism and/or an encouragement thought.

References

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies, 4*(3), 114-123.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An Empirical Evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction, 24*(6), 574-594. doi:10.1080/10447310802205776

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. doi:10.1080/10447318.2015.1064648

Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the System Usability Scale: a test of alternative measurement models. *Cognitive processing, 10*(3), 193-197.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry, 189*(194), 4-7.

Dawe, S. R., Windsor, J. A., Broeders, J. A., Cregan, P. C., Hewett, P. J., & Maddern, G. J. (2014). A systematic review of surgical skills transfer after simulation-based training: laparoscopic cholecystectomy and endoscopy. *Annals of surgery, 259*(2), 236-248.

Digital Surgery Limited, D. (n.d., 30.07.2020). Touch Surgery. Retrieved from https://play.google.com/store/apps/details?id=com.touchsurgery

Gallagher, A. G., Leonard, G., & Traynor, O. J. (2009). Role and feasibility of psychomotor and dexterity testing in selection for surgical training. *ANZ Journal of Surgery, 79*(3), 108-113. doi:10.1111/j.1445-2197.2008.04824.x

Graafland, M., Schraagen, J. M., & Schijven, M. P. (2012). Systematic review of serious games for medical education and surgical skills training. *British journal of surgery, 99*(10), 1322-1330.

Greco, E. F., Regehr, G., & Okrainec, A. (2010). Identifying and Classifying Problem Areas in Laparoscopic Skills Acquisition: Can Simulators Help? *Academic Medicine, 85*(10), S5-S8. doi:10.1097/ACM.0b013e3181ed4107

Hofstad, E. F., Våpenstad, C., Chmarra, M. K., Langø, T., Kuhry, E., & Mårvik, R. (2013). A study of psychomotor skills in minimally invasive surgery: What differentiates expert and nonexpert performance. *Surgical Endoscopy, 27*(3), 854-863. doi:10.1007/s00464-012-2524-9

International Organization for Standardization. (2018). *Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts* (ISO Standard no. 9241-11). Retrieved from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en

Kaya, A., Ozturk, R., & Gumussoy, C. A. (2019). Usability measurement of mobile applications with system usability scale (SUS). In *Industrial Engineering in the Big Data Era* (pp. 389-400): Springer.

Kneebone, R. (2010). Simulation, safety and surgery. *BMJ Quality & Safety, 19*(Suppl 3), i47-i52.

Kowalewski, K.-F., Hendrie, J. D., Schmidt, M. W., Proctor, T., Paul, S., Garrow, C. R., . . . Nickel, F. (2017). Validation of the mobile serious game application Touch Surgery™ for cognitive training and assessment of laparoscopic cholecystectomy. *Surgical Endoscopy, 31*(10), 4058-4066.

Krol, M. W., de Boer, D., Delnoij, D. M., & Rademakers, J. J. (2015). The Net Promoter Score–an asset to patient experience surveys? *Health Expectations, 18*(6), 3099-3109.

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. (2009). *The factor structure of the system usability scale.* Paper presented at the International conference on human centered design.

Lewis, J. R., & Sauro, J. (2017). Revisiting the Factor Structure of the System Usability Scale. *Journal of usability studies, 12*(4).

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring perceived usability: The SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction, 31*(8), 496-505.

Petrie, H., & Bevan, N. (2009). The Evaluation of Accessibility, Usability, and User Experience. *The universal access handbook, 1*, 1-16.

Pierorazio, P. M., & Allaf, M. E. (2009). Minimally invasive surgical training: Challenges and solutions. *Urologic Oncology: Seminars and Original Investigations, 27*(2), 208-213. doi:10.1016/j.urolonc.2008.09.017

Royal Australasian College of Surgeons, R. (n.d., 28.04.2018). SimuSurg. Retrieved from https://play.google.com/store/apps/details?id=com.cmee4.endoapp

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research*: Morgan Kaufmann.

Van Der Poel, H., Brinkman, W., Van Cleynenbreugel, B., Kallidonis, P., Stolzenburg, J. U., Liatsikos, E., . . . Dasgupta, P. (2016). Training in minimally invasive surgery in urology: European Association of Urology/International Consultation of Urological Diseases consultation. *BJU International, 117*(3), 515-530. doi:10.1111/bju.13320

Appendix

Appendix A. The informed consent, with the three components

Invitation letter

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

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.

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;

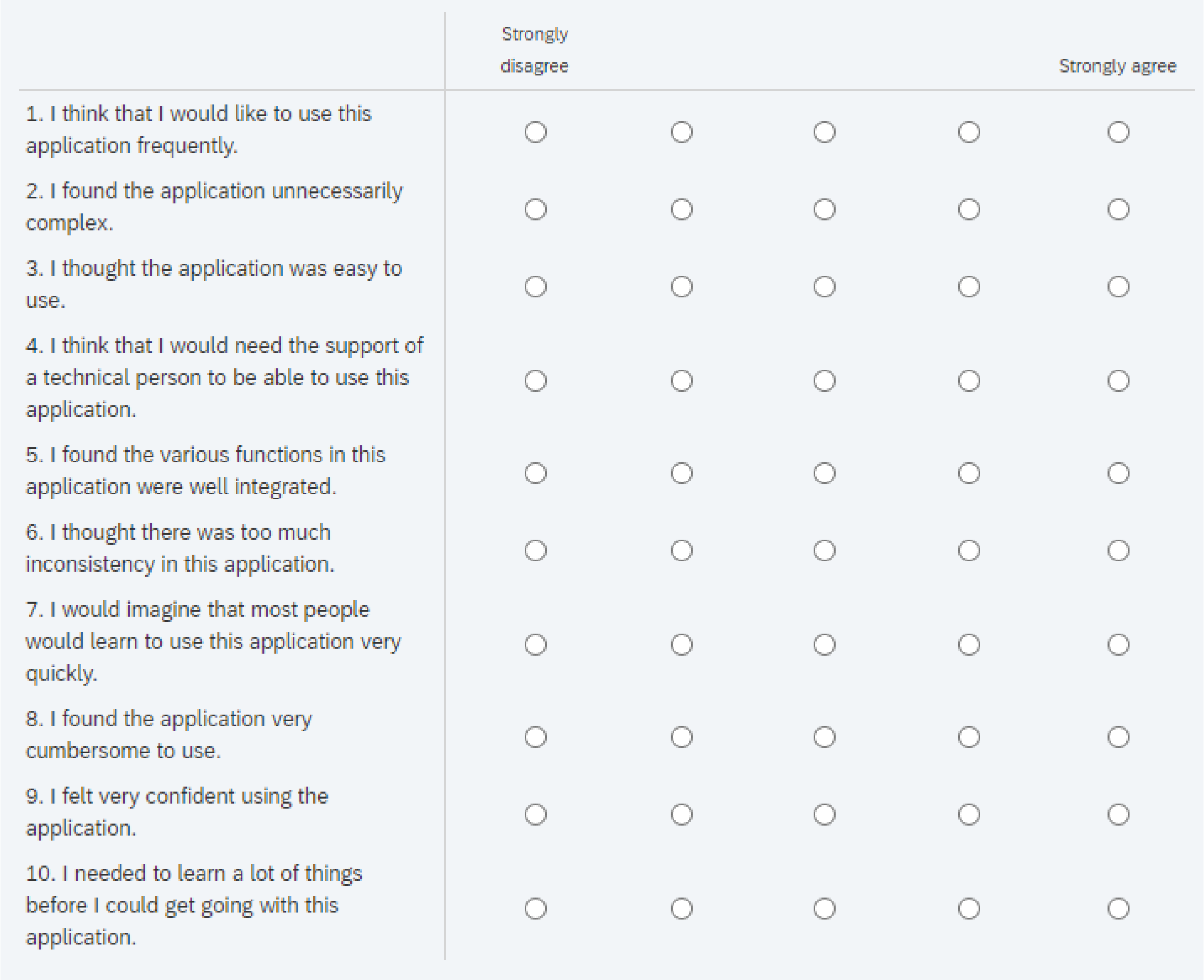
- The applications I am going to use might also make use of some of the information I provide (e.g. results of the simulation).

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 B. The version of the SUS used in this study



	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.	○	○	○	○	○

Appendix C. 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.

Task: 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 only 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 do not 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 a.amariei@student.utwente.nl 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 D. Inconclusive results

The dimensionality of Touch Surgery

When not imposing a number of factors, 4 components were extracted for the *novice* group, 3 components for the *intermediate* group and 3 components for the *expert* group. With the 2 component structure forced, the bi-dimensionality nature was visible in the *novice* group, but not in the other two. Varimax rotation was used for all analyses.

The intermediate group

With the forced 2 component structure, the components of the SUS were not clearly represented. Factors loaded in an unexpected manner in both components. Based on the previous KMO and Bartlett's Test results for this group, it was safe to consider the *intermediate* sample not representative for the target group.

Table with the factor loadings of the two-components across all groups

Table 4

Factor loadings across all groups

Group	Novices		Intermediates		Experts	
Items	Usability	Learnability	Usability	Learnability	Usability	Learnability
1	**.739**	.188	.256	**.627**	**.645**	-.084
2	**.700**	.353	**.731**	-.278	**.756**	-.012
3	**.831**	.212	**.428**	-.405	**.456**	**.506**
4	.138	**.769**	**.767**	**.428**	.016	**.723**
5	**.811**	-.093	.326	**.834**	**.670**	**.592**
6	**.577**	.100	**.865**	-.292	**.652**	**.627**
7	**.523**	**.613**	**.833**	-.227	.341	**.795**
8	**.787**	.122	**.739**	-.038	**.777**	.394
9	**.733**	**.474**	**.701**	.014	**.790**	**.413**
10	.009	**.807**	**.898**	.086	-0.33	**.922**

Note. Loading greater than .4 are in bold.