Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

POCUS – Computer-based Training for Improving the Quality of Ultrasonic Findings in Gallbladder Changes

MASTER-ARBEIT

2022

Autorin
Lehmann, Manuela

Begleitperson
Michel, Stefan

Praxispartnerin
CASRA
Hardmeier, Diana

Abstract

Point-of-care-ultrasound (POCUS) gains attention and is increasingly used in the medical community as it is applied by attending physicians. However, the lack of training in visual search in medicine put the success of POCUS at risk. The medical community requires standardized training. An attempt at that is shown in the presented study. Medical students (n = 14) who conducted an oral presentation as a passive learning method, followed by a computer-based training as an active learning method could increase their proportion of correct responses for ultrasound images of the gallbladder significantly compared to students who only conducted the passive learning method (n = 16) or no training at all (n = 16). Participants also mentioned their interest in an online training tool for their own universities. The successful application of POCUS would increase patient safety and lowers medical costs, but sufficient training is needed to fulfill these advantages. This paper contains 97’023 characters (incl. spaces, without appendices).

Keywords: Point-of-care-ultrasound (POCUS); visual search; computer-based training (CBT); passive and active learning method; gallbladder alterations; ultrasound

Zusammenfassung

Point-of-Care-Ultrasound (POCUS), eine eher neuere Untersuchungsmethode für behandelnde Ärzte direkt am Patientenbett, erfährt momentan einen Aufschwung in der Medizin. Die Tatsache, dass die visuelle Suchaufgabe im medizinischen Bereich jedoch zu wenig an den Universitäten gelehrt wird, gefährdet den Erfolg von POCUS. Die vorliegende Studie versucht einen Ansatz für ein standardisiertes Training anzustreben, so wie es von vielen Medizinern und Medizinerinnen gefordert wird. Medizinstudierenden (n = 14), welche eine Kombination von einer passiven Lerneinheit (vertonte Präsentation) und einer aktiven Lerneinheit (computer-basiertes Training) absolvierten, konnten ihren Anteil in korrekten Antworten in Ultraschallbildern der Gallenblase erfolgreich erhöhen, im Gegensatz zu den Medizinstudierenden (n = 16), die nur die passive Lerneinheit absolviert hatten oder gegenüber der Gruppe (n = 16), die kein Training durchführte. Der richtige Einsatz von POCUS verspricht eine erhöhte Patientensicherheit und tiefere Behandlungskosten, jedoch braucht es ein Training für POCUS, um diese Vorteile erreichen zu können. Die vorliegende Arbeit enthält 97’023 Zeichen (inkl. Leerzeichen, ohne Anhang).

Schlüsselwörter: Point-of-Care-Ultrasound (POCUS); Visuelle Suche; Computer-basiertes Training (CBT); aktive und passive Lernmethoden, Veränderungen der Gallenblase; Ultraschall

Table of Content

1 Introduction

Visual search, which describes the task of searching for a target within a scene surrounded by distractors, is inherent in our everyday life (Carrigan et al., 2015; Yang et al., 2002). To understand a scene, we fixate our eyes on specific areas in the scene and move between them, rather than moving the eyes randomly. The *Saccadic* is the most important to consider in visual search among the most common eye movements. *Saccadic* is the voluntary eye movement where the foveal vision, which is the best visual acuity, is directed to a specific point of interest. During a visual search, the area of interest usually needs examination for longer than 100 milliseconds for the brain to register that area (Yang et al., 2002). However, visual search is not only used in everyday life, but it also finds its place in fields that are considered "high consequences". That is because of their potentially high costs of inspection errors such as injury, mortality, loss of expensive equipment, scrapped items, rework, or failure to procure repeat business. Next to the field of nuclear weapons, nuclear power, airport baggage screening, aircraft maintenance, and the food industry, medicine is also considered a high-consequence field (See et al., 2017). The visual search task in medicine is considered difficult, even for well-trained experts, and even if the target is visible for an extended amount of time (Wolfe, 2010). Wolfe (2010) also implied that technology is needed to support humans and eliminate errors in this challenging visual search task. Until technology can fulfill this task, it is necessary to understand how a target in visual inspection can be missed or how an image can be declared as containing a target even though untrue, hence a false alarm. For example, a miss in cancer screening which leads to cancer death, is more expensive than false alarms. However, the false alarm also comes with financial and emotional costs, demonstrating the importance of visual search tasks in medicine (Carrigan et al., 2015; Wolfe, 2010).

Kundel, Nodine, and Carmody (1978) state that three types of medical imaging errors can occur. The visual search error, where the examiner never sees the abnormality (30 %), the recognition errors, where the abnormality is hidden, and therefore no meaning is given to the area even though the abnormality was fixated (25 %), and the decision error, where the abnormality is fixated but not declared as abnormality (45%). Due to the trend of medical imaging practices toward higher volume and increasingly complex examination, human error will also possibly increase, leading to a 30 % miss and an equally high false alarm rate (Berlin, 2005; Carrigan et al., 2015).

Medical imaging technologies come in different forms, like x-rays, fluoroscopy, ultrasound, computer tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET), and have become an inherent part of modern healthcare (Yang et al., 2002). Radiologists who examine the medical images are experts in visual search in medicine. The task involves examining a cluttered medical image followed by a diagnostic decision based on abstract anatomical features. The radiologists need to be highly accurate in their performance whose level of expertise can only be reached by comprehensive training and practice (Carrigan et al., 2015). A study by Gunn, Jones, Bridge, Rowntree, and Nissen (2018) has shown significant results regarding first-year medical imaging students who conducted two challenging procedures for novices (posterio-anterior scaphoid and dorsi-plantar oblique foot) one in the traditional practical laboratory experience and the other using a VR simulation software. In an afterward conducted role-play task of these two procedures, the students trained using the VR software could improve their skill level significantly to 4.75 %. This study demonstrates that realistic VR training can enhance the obtainment of technical skills of medical imaging students. Additionally, the implementation of this software could show key findings of student enjoyment and independent learning in a

previously conducted study (Bridge et al., 2014). The beneficial effect of early exposure to didactic teaching in radiology, which resulted in higher acknowledgment of the importance of radiology to the general practice of medicine and better score in a basic radiologic knowledge test in first-year medical students, was shown in a study by Branstetter, Faix, Humphrey, and Schumann (2007). Next to the radiologists, another medical occupation needs to be taken into account. In Australia, sonographers perform diagnostic scans and select images before reporting to the radiologist. Therefore, abnormalities that the sonographer does not detect cannot get reported to the radiologist, which could lead to missing the abnormalities. Even though the underlying visual tasks and the potential for errors are similar to those of radiologists, the work of sonographers also contains other aspects. The task of sonographers is unique considering there being continuous visual and cognitive tasks, making judgments, and capturing pathologies in real-time. Therefore the tasks of sonographers should be further examined (Carrigan et al., 2015). In a study by Mendiratta-Lala, Williams, de Quadros, Bonnett, and Mendiratta (2010), 29 radiology residents with four years of training were tested in their written, practical, and technical abilities considering an ultrasound-guided procedure. The pretest contained a 20-question test and a web-based module that addressed the fundamental indications, contradictions, and techniques to perform an ultrasound-guided procedure. The web-based module included a teaching video of 10 minutes, and the posttest also contained a simulated procedure. Afterward, the residents received training of staff body interventional radiologists, answered a questionnaire of six questions, and were instructed to train in the following six months at their convenience before the residents conducted the posttest. The results showed a significant increase from the pretest to the posttest in their written tests ($p = 0.024$) and their practical scores ($p < 0.001$), while a trend could be seen in their improvement in procedural-based skills

($p = 0.07$). Mendiratta-Lala, et al. (2010) state that the results of training with simulation can increase patient safety. The training includes the different rates of learning of the trainees, which may be a better measure of the actual skill level. Furthermore, it allows standardization of the training, lets different teachers and their different teaching methods disappear, and opens the possibility of lower cost than the conventional training.

Ultrasound has gained attention in the last few years because of POCUS, which means "Point-of-Care Ultrasound". POCUS is a method of ultrasound that is used at the bedside and covers the examination of the entirety of the patient's body, including examinations in the fields of cardiology, critical care medicine, emergency medicine, pulmonary medicine, and many more (Moore & Copel, 2011; Osterwalder & Tercanli, 2018). This new method of POCUS is meant to expand the clinical examination. It provides the attending physician, who must not be an expert in ultrasound, with an additional tool for making reliable and optimal decisions and monitoring the patient. POCUS, for example, can reduce the time establishing a diagnosis, the number of needed CT examinations, and costs (Osterwalder & Tercanli, 2018). Furthermore, POCUS can improve patient care by providing real-time, non-invasive, non-radiating, and low-cost imaging to help guide clinical decision-making and is easy to repeat if the patient's condition changes (Moore & Copel, 2011). Unlike concerns about medical imaging technologies like the ionizing radiation from computer tomographic (CT) scanning, which is increasingly recognized as a potential cause of cancer, ultrasonography has not shown any epidemiologic evidence of harmful effects at normal diagnostic levels even though it has been used for decades (Barnett, 2002; Brenner & Hall, 2007). The possibility of attending physicians carrying and using a personalized ultrasound device will have major effects on the clinical routine and ultimately replace

the stethoscope (Hossfeld, 2020; Osterwalder & Tercanli, 2018). Since ultrasound, in general, is highly examiner-dependent and as POCUS will be used by any kind of attending physician, it is on the bottom of an examiner-expert pyramid, which will consequently lower the quality of ultrasound. Therefore, POCUS is met with great concerns within the medical community (Osterwalder & Tercanli, 2018). How POCUS can be taught and its effectiveness is demonstrated in multiple studies (see Alba et al., 2013; Greenstein et al., 2017; Maw et al., 2016). Yamada et al. (2018) stated that since POCUS is a relatively new method, there are young medical students or residents who want to learn POCUS and experienced physicians who want to start using this newer method. Their study investigated what kind of training was needed to teach novice trainees who were medical students and residents in their first five postgraduate years and novice attending physicians who were physicians in their sixth or higher postgraduate year. Furthermore, they wanted to explore if the same kind of training could be used to educate the two different learning groups because of the different amount of previous medical experience. The participants were taught in a one-day POCUS course that included internet-based learning, lectures, and hands-on sessions. A written test and a self-evaluation exercise were conducted at the beginning and at the end of the one-day course in order to investigate the change in their POCUS skills. The study could show that the novice trainees and the novice attending physicians could both improve their POCUS knowledge, image interpretation skills, and confidence in their skills significantly. Stating that a one-day course is already enough to have significant improvements, the authors challenge the amount of training needed for significant improvements in POCUS skills for further research. Finally, they concluded that standardized POCUS training curricula are needed throughout Japan (Yamada et al., 2018). Through training, the confidence in the own skills increases, as seen in the

study of Yamada et al. (2018). However, overconfidence describes that one overestimates the accuracy of the own judgment or decision. Even though overconfidence has advantages, like being esteemed by their peer and allowing people to escape stress associated with pessimistic thoughts, it also suppresses the delight of success and shows most trouble in its potential cost (Anderson et al., 2012; Armor & Taylor, 1998; Dunning & Griffin, 1990; McGraw et al., 2004). In one of their studies, Sanchez and Dunning (2018) showed that a "beginner's bubble" occurs after a small amount of training. The participants overestimated their abilities to an unacceptable level, even though they had no prior knowledge of the tested material. However, the overconfidence deflates to a normal confidence level with further training. This finding is also supported by a study where medical students (fourth year), medical residents (second and third-year internal medicine), and general internists were tested on their confidence and accuracy in stating a diagnosis. Medical students showed the least accuracy and confidence. General internists had the highest accuracy and confidence. Surprisingly, the medical residents showed more confidence about the correctness of their diagnoses but were less accurate than attending physicians (Friedman et al., 2005). Therefore, the presented study will measure the participants' confidence in detecting diseases or abnormalities.

Osterwalder and Tercanli (2018) state that comprehensive ultrasound performed by a specialist is clearly of superior quality and can therefore not be replaced by POCUS. Nevertheless, an examination through an ultrasound performed by an expert is not always available due to insufficient time. Therefore, in combining the advantages of the ultrasound performed by experts and POCUS, which is used by attending physicians, Osterwalder and Tercanli (2018, p. 607) define six measures that have to be met:

(1) The concept of the conventional US learning pyramid must allow room for recognition of the similarities and differences between the methods and the fact that POCUS represents a new readily available but limited ultrasound method.
(2) Standardized, evidence-based training curricula and standards for POCUS must be created and various expert levels for training must be defined.
(3) Tutors need to be trained and training centers for those interested in POCUS must be provided.
(4) Ultrasound experts need to develop guidelines as to when POCUS patients would profit from a referral to a specialist and should be involved in training.
(5) It must be recognized that POCUS is not easier to perform even though POCUS examinations are faster and limited in their scope. The POCUS spectrum ranges from simple to complicated.
(6) POCUS examiners must undergo good training and supervision, observe the POCUS principles, and avoid overinterpretation and misinterpretation.

The following study grounded its idea in regards to the need for training that physicians in all fields of expertise and experience in medicine can perform. Another part of this presented study considers the already known theory of active and passive learning. Passive learning is described as the traditional way for professors to deliver lectures to a great number of students simultaneously with little opportunity for the students to actively engage in the lecture, for example, through discussions or experimental exercises (Wingfield & Black, 2005). While the advantage of the passive teaching method is that the professor can present a large amount of material in a relatively short amount of time, the disadvantages relate more to the students having problems retaining this amount of knowledge and lacking attention during the lecture (Dorestani, 2005; Miner et al., 1984; Van Eynde & Spencer, 1988). Compared to that,

active learning involves the students both physically and mentally (Bonwell & Eison, 1991). Active learning methods describe practices where the students are actively engaged, such as small-group discussions, short writing exercises, field trips, role play, or student self-assessment exercises (Bonwell & Eison, 1991; Ebert-May et al., 1997; Sarason & Banbury, 2004). The advantages of the active learning methods are that the students are more engaged in reading, writing, and discussions during a lecture. Their motivation is increased, and the students can receive immediate feedback. The students are also more engaged in higher-order thinking, such as analysis, synthesis, and evaluation (Bonwell & Eison, 1991). Studies by Dorestani (2005), Sarason and Banbury (2004), and Benek-Rivera and Matthews (2004) show that active learning is more effective than passive learning. However, Michel, Cater, and Verela (2009) state that barely any study compares the two methods quantitatively. They rather focus on attitudinal reactions of the students than cognitive outcomes. Additionally, the broad range of activities described as active learning complicates this field of research even more. However, McDonald and Frank (2016) argue that passive and active learning must not be viewed as two separately used methods and at the same time cannot be used to complement each other. Their study shows that a previous passive learning task increases the exploration of the followed active learning task and shows overall a better performance in an abstract concept learning task. Even though a combination of two active learning tasks showed slightly better results, McDonald and Frank (2016) state that the succession of passive and after active learning tasks is beneficial due to the reduced learning cost. Nevertheless, active learning is also used in the medical field. Graffam (2007) describes active learning methods such as “breaks as action moments”, “questioning techniques”, and “case scenarios” as starting points for medical faculties to engage in the promising application of active learning. This pedagogical change is

desperately needed in medical education because medical schoolteachers usually lack pedagogical training. No change can be seen in teaching medicine even though the science of medicine changed tremendously (Cohen, 2004; Hurst, 2004).

The following presented study is based on the not yet published study of Lehmann & Michel (2020), which is further explained in 2.3.3 and 2.3.4. Due to the significant findings in this not published study, two ultrasound experts were interested in this presented study and agreed to support it from a medical perspective. The presented study designed training and a test for medical students, which focuses on visual search in ultrasound images of the gallbladder. The training contains two parts. One is an oral presentation about diseases and abnormalities of the gallbladder seen in ultrasound images where the participants just listen and therefore is considered a passive learning method. In contrast, the second part is a computer-based training. It is considered an active learning method due to the active interaction with the training software, which presents the ultrasound images and the immediate feedback to the user on their chosen decision and further explanation of the shown disease or abnormality. The following research question was investigated in the presented study:

What differences can be found between a passive and the combination of a passive and active training session in terms of detection performance of gallbladder ultrasound images in medical students?

As described in the literature, the active learning method should lead to better results in the detection performance of the students but is even more increased by the combination of passive and active, as shown by McDonald and Frank (2016). The test considers a wide range of gallbladder diseases and abnormalities and does not focus on just one disease present in the images. Next to the investigation of the detection performance, another important factor is how confident the attending physicians are in

their decision. Since the lack of confidence is considered weak and patients usually value confidence over uncertainty, the importance of confidence is clearly stated (Croskerry & Norman, 2008). Therefore, it is also part of the presented study. Since POCUS is also used in emergency medicine where time is critical, this study's participants' response time is measured and analyzed as well (Fatovich, 2002). Regarding the aspect of time and speed, the occurrence of a speed-accuracy-trade-off must be mentioned. The speed-accuracy-trade-off is an aspect of the decision-making process and states that speed and accuracy often stand against each other. In a task, where time is more relevant than accuracy, a person will focus on speed and may lose accuracy instead. In contrast, decisions are slower when a task emphasizes accuracy (Bogacz et al., 2010; Franks et al., 2003). Favored in the presented study is that no speed-accuracy-trade-off can be found because, through training, the participants should increase their detection performance as well as their speed.

The following hypotheses were created along with the presented literature and are being used as guidelines in the presented study:

Detection Performance	Detection performance will increase significantly by combining the passive training session (oral presentation) and the active training session (computer-based training).
Confidence Rating	Confidence about the given response will increase significantly by combining the passive training session (oral presentation) and the active training session (computer-based training).
Reaction Time	Participants will need significantly less time to investigate ultrasound images by combining the

> passive training session (oral presentation) and the active training session (computer-based training).

If this presented study shows significant results benefitting the computer-based training combined with the oral presentation, it would be possible to create standardized training for different human body areas examined by POCUS. Additionally, a standardized test could be created to certify medical students, medical residents, or doctors interested in POCUS, which would meet some of the measures introduced by Osterwalder and Tercanli (2018).

2 Methods

This chapter describes the different methods of the presented study. First, the group of participants will be described, followed by the procedure of the study. Afterward, the used materials will be explained, including the two questionnaires, both oral presentations, the test (M-UCAT), the training (M-UTT), and at last, the used statistical analyses.

2.1 Participants

In this study, 59 students of medicine initially participated voluntarily after receiving a study description (Appendix A). However, 13 participants did not finish the study due to various personal reasons. After their exclusion, 46 students of medicine participated, whereas 15 participants were male (33 %) and 31 females (67 %) between the age of 20 – 31 (M = 24,33, SD = 3,09). After a questionnaire about their demographics (see 2.3.1) and the first Medical Ultrasound Competence Assessment Test (M-UCAT) (see 2.3.3), three groups were formed, two experimental groups and a control group. These three groups were as similar as possible based on the questionnaire

results about their demographics and the first M-UCAT. Since only one defining variable was normally distributed in all three groups, a Kruskal-Wallis-H-Test was used to ensure no significant differences existed between the three groups. The following variables were regarded: proportion of correct answers from the first M-UCAT, confidence rating, reaction time, age, and sex (all p-values are > .216) (Appendix B). Both experimental groups participated in the training, whereas the control group did not conduct any exercise.

2.2 Procedure

A graphic presentation of the study design can be seen in **Figure 1**. The participants were contacted by e-mail after expressing their interest in participating in the study. First, they received an e-mail explaining the next steps in the study and the link to the questionnaire about their demographics (see 2.3.1). By pressing on the link, the participants were directed to the website where the questionnaire took place, and the participants could navigate through the questions with their mouse or trackpad. After all participants had finished the questionnaire, they received a link to an oral presentation (see 2.3.2) about the basic information of sonography and how it is used correctly. This oral presentation was created by an expert in sonography and medical education. The study's next step was for the participants to complete the first M-UCAT. By e-mail, the participants received their login, information about the M-UCAT, and a document containing the written-out names of the diseases because, due to the software setting, some of the disease's names had to be shortened in the M-UCAT. After the participants logged in to the software, some more information was shown about the interface of the M-UCAT, how the software must be handled, and what they could expect in the actual test. Before the actual test started, four trial images were shown, and the participants could get used to the software and prepare for the upcoming test. Subsequently, the

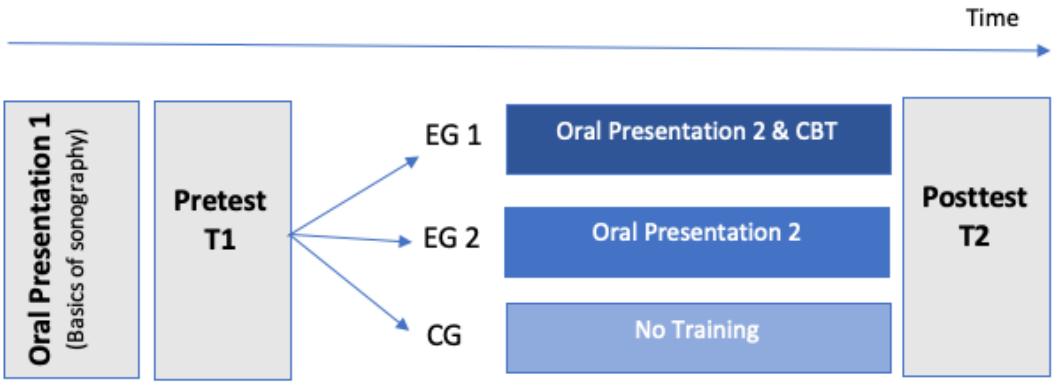
actual test started, and the participants navigated through with their mouse or trackpad. According to the questionnaire data and the first M-UCAT, two experimental groups and one control group were formed. Experimental group 1 (n = 14) listened to a second oral presentation about abnormalities and diseases of the gallbladder (see 2.3.2) and performed a computer-based training with the Medical Ultrasound Tutor training system (M-UTT) (see 2.3.4). The second oral presentation was prepared by a leading expert in ultrasound and emergency medicine and could be reached by the participants through a link sent by e-mail. The participants of experimental group 1 were instructed to listen to the oral presentation first and then complete the M-UTT. Therefore, the participants had to confirm by e-mail that they had listened to the oral presentation and only got access to the M-UTT afterward. For the M-UTT, the participants could log into the same software as for the M-UCAT and got more information about the M-UTT by e-mail as well as the document with the written-out names of the diseases used in the M-UTT. Before starting the training, four trial images were shown to allow the participants to get used to the training software. Experimental group 2 (n = 16) was instructed to only listen to the second oral presentation about abnormalities and diseases of the gallbladder. In contrast, the control group (n= 16), group 3, did not engage in any kind of training. After the pieces of training were completed, all participants conducted the second M-UCAT by using the same login as in the first M-UCAT. To give the participants of experimental group 1 the opportunity to express their opinion about the computer-based training and possible improvements, a short questionnaire (see 2.3.1) was conducted at the end, which could be entered by a link sent by e-mail.

All questionnaires, tests, and pieces of training were conducted online. The participants used their own computer and additional equipment (e.g., keyboard and

mouse). The communication between the participants and the study team was carried out by e-mail.

Figure 1

Study Design



Note. EG is shortened for Experimental group while CG stands for control group

2.3 Materials

In this chapter, the materials used in this study are explained. The order follows as much as possible the procedure of the study. Therefore, it begins with the questionnaires, followed by the oral presentations, the Medical Ultrasound Competence Assessment Test (M-UCAT), and the Medical Ultrasound Tutor training system (M-UTT).

2.3.1 Questionnaires

The first questionnaire collected the needed demographic data to create three statistically equivalent groups: experimental group 1, experimental group 2, and the control group. This questionnaire was sent to all participants who agreed to participate in this study before listening to the first oral presentation. The questionnaire collected data about the participant's name, age, sex, university, experience with ultrasound images, and experience with gallbladder abnormalities and diseases.

The second questionnaire was only sent to the experimental group 1 who had completed the M-UTT. This survey asks the participant about their opinion of the M-UTT, a general online-training tool for their study, and how x-ray image interpretation is taught at their universities.

Both surveys were created and conducted with Tivian EFS (survey.fhnw) and can be found in Appendix C and Appendix D.

2.3.2 Oral Presentations

The first oral presentation was used at the beginning of the study after the participants answered the questionnaire about their demographics and before the first M-UCAT was conducted. The oral presentation could be accessed through a link sent to the participants by e-mail. The topic of this first oral presentation covered the basics of sonography and how it is used correctly. More specifically, the first oral presentation included information about where sonography is used, the different frequencies of the different ultrasound methods, how ultrasound works in medicine, and how the images must be interpreted. The first oral presentation runs 32 minutes and 43 seconds. The screenshots of this first oral presentation can be seen in Appendix E.

The second oral presentation is used for experimental groups 1 and 2. For experimental group 1, it is meant to be used before the participants conduct the computer-based training (M-UTT), whereas, for experimental group 2, the second oral presentation presents the entirety of the training. The topic of the second oral presentation is the gallbladder. It includes information about its anatomy, how it is examined with ultrasound, and usual errors while interpreting ultrasound images of the gallbladder. The second presentation runs 29 minutes and 5 seconds, and the screenshots of the presentation can be seen in Appendix F.

2.3.3 Medical Ultrasound Competence Assessment Test (M-UCAT)

The M-UCAT is based on the X-Ray CAT, which was initially developed in aviation security (*CASRA*, n.d.). The X-Ray CAT measures the x-ray image interpretation competency of airport security screeners under consideration of the Signal Detection Theory of Green and Swets (1966). This theory describes the calculations of a detection performance (d' and A') considering the four possible responses when searching for a signal (target object) surrounded by noise (everything else in the image). The four possible responses are shown in **Figure 2**.

Figure 2

Responses according to the Signal Detection Theory (Green & Swets, 1966)

		Signal	
		present	Not present
Response	NOT OK	Hit	False Alarm
	OK	Miss	Correct Rejection

Note. Own representation of the four responses considered in the calculation of the detection performance (d' or A') in Signal Detection Theory (Green & Swets, 1966). The response "NOT OK" is used when it is believed a target object is present, whereas the response "OK" is used when it is believed that no target object can be seen in the image

In a not yet published study (Lehmann & Michel, 2020), the software and the basic knowledge of the X-Ray CAT were used to create a new Competence Assessment Test in the field of medicine. This test contained x-ray images of the thorax where the

lungs were shown. The participants had to decide whether a pleural effusion, which is an excessive accumulation of fluid between the lung and chest tissue (Karkhanis & Joshi, 2012), was present or not. Based on this study, where the participants could significantly increase their detection performance, the M-UCAT was created.

The M-UCAT contains ultrasound images of the gallbladder. Two doctors who acted as medical consultants for the presented study and their network of doctor colleagues provided these images. One of the two mentioned doctors also created the oral presentation about the abnormalities and diseases of the gallbladder. The medical professionals also marked the diseases and abnormalities on the images.

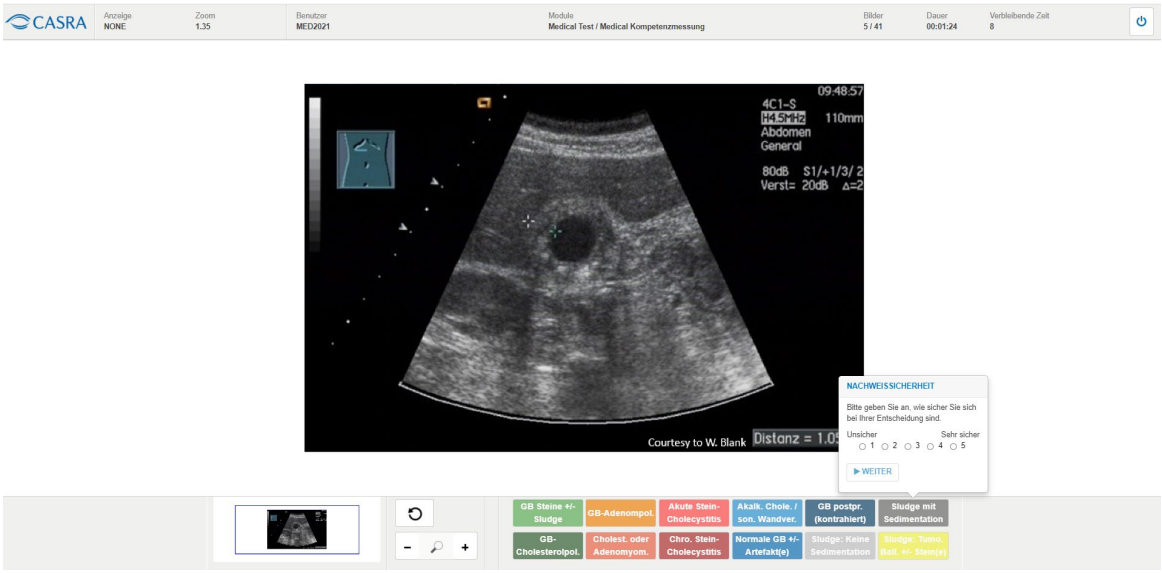
Before the participants could work on the ultrasound images, the M-UCAT provided preliminary information about the test interface, the task the participants had to conduct, and the possible responses (Appendix G). Afterward, the participants could work on four practice images to get used to the test interface and prepare themselves for the upcoming test. While working on the practice images, the participants got feedback on whether their chosen answer was correct or not and some additional information about the image. This feedback should help the participants get used to the abnormalities and diseases, but the feedback was not included in the actual test, so no training was possible. The actual test contained 41 images, and the following abnormalities and diseases of the gallbladder could be seen on the images next to images of normal gallbladders:

Gallstones with or without sludge (5 images)	Acalculous cholecystitis or other wall thickening (edema, etc.) (3 images)
Cholesterol polyp (3 images)	Normal gallbladder with or without artifacts (4 images)
Adenompolyp (7 images)	Gallbladder postprandial (contracted) (3 images)
Cholesterolosis or Adenomyomatosis (2 images)	Sludge: No sedimentation (3 images)
Acute cholecystitis with gallstones (4 images)	Sludge with sedimentation (1 image)
Chronic stone cholecystitis with gallstones (5 images)	Sludge: Tumorous cluster with or without gallstones (1 image)

The participants had to decide which of the 12 possible answers could be seen on the ultrasound image by clicking on the corresponding button for each image. Each image was shown for 20 seconds before it disappeared. If the participant did not choose an answer during that time, they had to decide on an answer without the visual help of the image. After each decision, the participant was asked to rate their confidence about their decision on a rating scale from “unsure” to “very sure”, where one of five options could be chosen. A screenshot of the M-UCAT with the confidence rating can be seen in **Figure 3**. The M-UCAT was planned to take around 15 to 20 minutes and ended with a thank you note as well as information about the next steps of the study.

Figure 3

Screenshot of the M-UCAT



2.3.4 Medical Ultrasound Tutor training system (M-UTT)

Alongside the X-ray CAT, which was used as the base for the M-UCAT, the M-UTT is based on the X-Ray Tutor training system, which was developed and still used to train security screeners in aviation security (*CASRA*, n.d.). This training software was also used in the not published study of Lehmann and Michel (2020). In the study, the participants trained for around 40 minutes for eight weeks to improve their detection performance on x-ray images of the lungs and the presence or absence of pleural effusion.

The images used for the M-UTT were also provided and assessed by the two medical consultants of this study and their medical network. The computer-based training contained a total of 61 ultrasound images of the gallbladder. It was separated into six blocks that dealt with a different kind of topic each, considering abnormalities and diseases of the gallbladder. None of the images used in the M-UTT was also used in

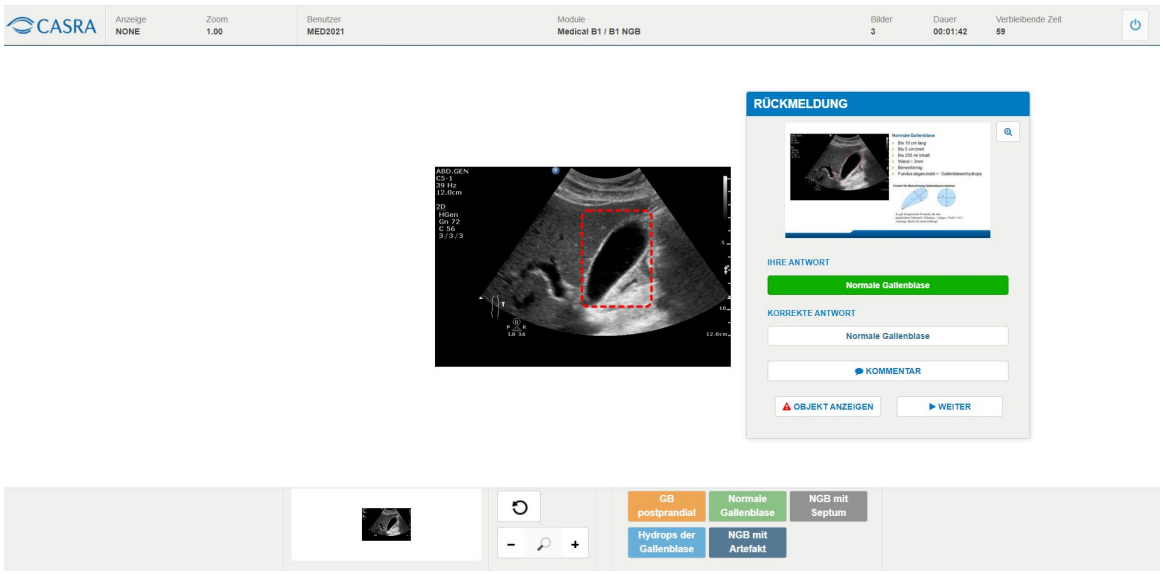
the M-UCAT. Afterward, there is an overview of the six computer-based training blocks and the number of images used.

Block number and topic	Possible responses	Number of images
Block 1: Variants normal gallbladder (13 images)	Gallbladder postprandial (contracted)	2
	Normal gallbladder	3
	Normal gallbladder with artifacts	3
	Normal gallbladder with septum	2
	Hydrops of the gallbladder	3
Block 2: Gallbladder stones (12 images)	Stone(s) with ultrasound shadow	6
	Stone(s) without ultrasound shadow	5
	Polyp(s)	1
Block 3: Sludge (8 images)	No sedimentation	3
	Sedimentation	3
	Tumorous cluster	2
Block 4: Gallbladder wall alterations (28 images)		
Block 4.1 Diffuse (14 images)	Acute cholecystitis with stone(s)	4
	Acute cholecystitis with stone(s) and hydrops	4
	Chronic cholecystitis with stone(s)	2
	Acalculous cholecystitis or wall edema	4
Block 4.2 Focal (11 images	Cholesterol polyp	3
	Adenoma polyp	5
	Adenomyomatosis	3
Block 4.3 Wall inclusions (3 images)	Cholesterolosis	2
	Normal gallbladder	1

Before every block of computer-based training, the participants got instructions about what topic the next block was about, what their task was, and how to exit the training if necessary (Appendix H). While working on the ultrasound images, the participants had to examine the respective image and decide which of the possible responses was correct. Then, by pressing the button with their chosen response, they triggered the feedback, that showed whether it was correct or not and provided some information about the image and its abnormalities and diseases or how a normal gallbladder can be detected if normal gallbladder was the correct answer. A screenshot of the M-UTT with the feedback can be seen in **Figure 4**. The images were shown for 60 seconds before they disappeared. If the participant did not decide on a response during that time, an answer had to be given by pressing the respected button without seeing the image anymore. To move on to the next block of computer-based training, the participants had to respond correctly 80% of the time. Until this passmark was met, the images repeated themselves after every image was shown at least once. Hence, the participants were able to use the gained information of the feedback about an image to examine the same image when it was shown repeatedly. The M-UTT was planned to take around 1 hour and 30 minutes. The participants could exit the training and continue later, but it was not defined by the study team when or for how long the participants train at once.

Figure 4

Screenshot of the M-UTT



2.4 Statistical Analyses

Under consideration that the Signal Detection Theory by Green & Swets (1966) builds the basis of the test and training system used in this study for the M-UCAT and M-UTT, it was initially planned to conclude the analysis by means of the Signal Detection Theory. However, after a first analysis and evaluation of the combination of the Signal Detection Theory and ultrasound images, it was unclear what the inherent measures of the theory (hit, correct rejection, miss, false alarm) mean in the field of medicine. Therefore, the analysis focused on the proportion of correct responses instead of the detection performance. All analyses, Kruskal-Wallis-H-Tests, Kolmogorov-Smirnov-Test, and Mann-Whitney-U-Test were conducted with SPSS version 27.0.1. All statistical tests employed an alpha significance level of .05. The effects are interpreted by the classification of Cohen (r = .10 equally a small effect; r = .30 equals a medium effect; r = .50 equals a strong effect.).

3 Results

This chapter is structured along with the three hypotheses investigated in this study. It starts with the proportion correct responses results, followed by the results of confidence rating, time of reaction, and lastly, the second questionnaire focusing on the opinion of experimental group 1 toward the conducted computer-based training. Proportion of correct responses, confidence rating, and reaction time were measured twice in this study with the M-UCAT, once at the beginning (pretest) and then again at the end (posttest). Between pre- and posttest, experimental groups 1 and 2 conducted the respectively assigned training. More precisely, experimental group 1 conducted the second oral presentation of approximately 29 minutes and the computer-based training where they trained for 54 minutes and 10 seconds (Min. 23 minutes 36 seconds / Max. 2 hours 12 minutes 32 seconds) on average. In addition, the participants of experimental group 1 needed, on average, 21 seconds to examine an image from the computer-based training. Experimental group 2 listened to the second oral presentation for approximately 29 minutes.

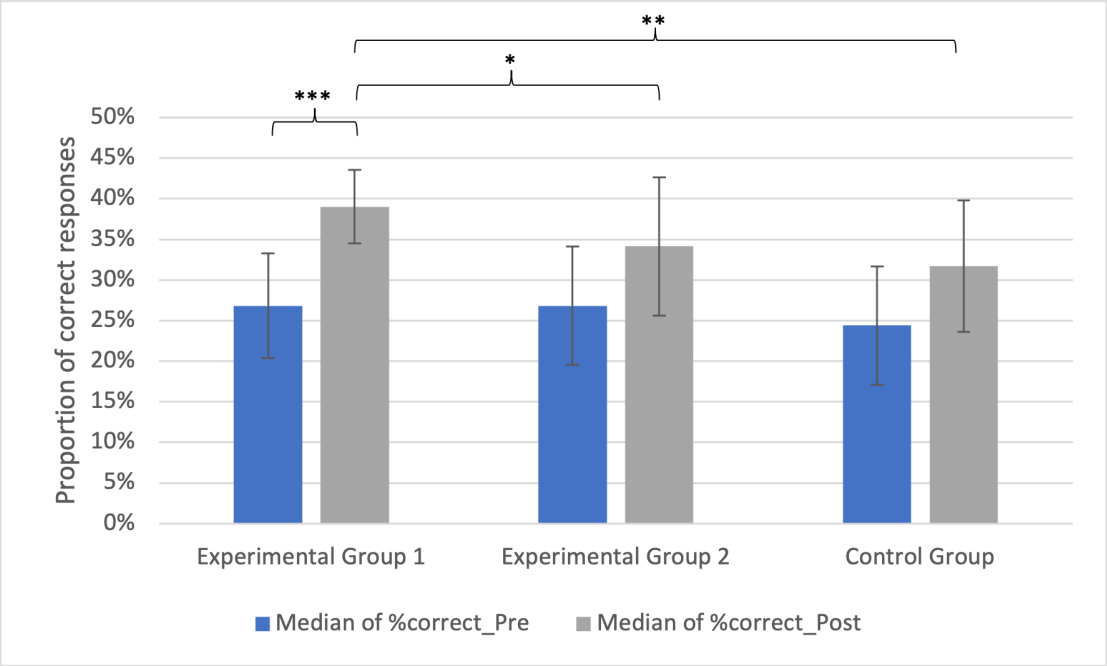
3.1 Correct Responses

As mentioned in chapter 2.4, the proportion of correct responses will be analyzed. The hypothesis for correct responses was that experimental group 1 would increase their ability of the proportion of correct responses significantly compared to experimental group 2 and the control group. At the beginning, a Kolmogorov–Smirnov test was calculated to determine if the variables are normally distributed. The proportion of correct answers in both pre- and posttest are not normally distributed (experimental group 1; pretest: $p = .200$, posttest: $p < .001$ / experimental group 2; pretest: $p = .130$, posttest: $p = .200$ / control group; pretest: $p = .061$, posttest: $p = .200$). Therefore, the

following analysis was executed with non-parametric methods. All Kolmogorov–Smirnov tests are documented in Appendix I.

Figure 5

Overview of Proportion of Correct Responses



Note. The median for each group regarding the proportion of correct responses in percent (y-axis). Error bars show the mean deviation from the median. *p < .05, **p < .01, ***p < .001.

Figure 5 shows the difference between experimental group 1, experimental group 2 and the control group considering the proportion of correct responses each for the pre- and posttest. A Paired-Sample Wilcoxon Signed Rank Test with the within-subject factor *test date* (pretest vs. posttest) was performed for all groups and visualized in violin plots (Appendix J). For experimental group 1 it revealed a large effect between the pretest (*Mdn* = 27 %) and the posttest (*Mdn* = 39 %) and can be seen in **Table 1**. No significant difference could be found between the pre- and posttest for experimental group 2 (pretest *Mdn* = 27 % / posttest *Mdn* = 34 %, *T* = 93, *z* = -1.298, *one-tailed p* = .102, *n* = 16, *r* = 0.32) and the control group (pretest *Mdn* = 24 % / posttest *Mdn* = 32

%, $T = 78$, $z = -0.518$, *one-tailed* $p = .311$, $n = 16$, $r = 0.13$). Boxplots containing the results from the pretest and posttest for all three groups can be seen in **Figure 6**.

Table 1

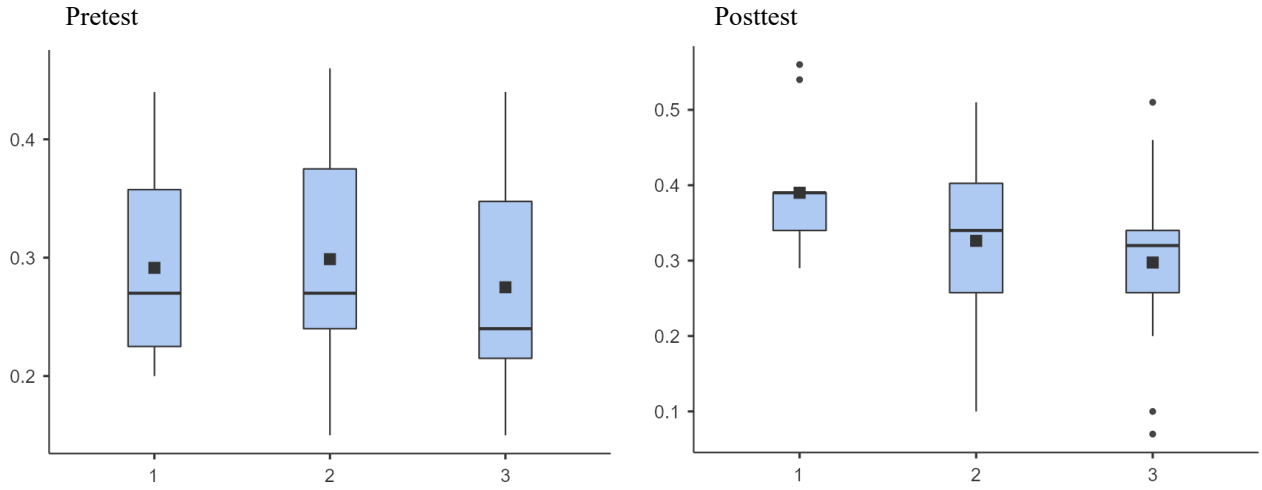
Paired-Sample Wilcoxon Signed Rank Test regarding the pretest and posttest of experimental group 1

	Proportion of correct responses of pretest		Proportion of correct responses of posttest					
Variable	*Mdn*	*MD*	*Mdn*	*MD*	*T*	*z*	*p (one-tailed)*	Cohen's *r*
Pre- and posttest for experimental group 1	27 %	6.4 %	39 %	4.5 %	102	-3.115	< .001	0.83

Note. Mdn = Median. MD = Mean deviation from the median. Experimental group 1 contains 14 participants.

Figure 6

Boxplots visualizing Proportion of Correct Responses Pretest and Posttest



Note. Boxplots of the pretest (left) and posttest (right) with proportion of correct responses in decimal (y-axis). The line represents the median while ■ represents the mean. Mild outliers are shown by● . 1 shows the results of experimental group 1, 2 shows the results of experimental group 2 and 3 shows the results of the control group.

Focusing on the pretest no significant differences could be revealed using a Kruskal-Wallis-Test between the experimental group 1, experimental group 2, and the control group (*Kruskal-Wallis-H (2)* = 0.986, *two-tailed* $p = .611$, $r = 0.15$).

Furthermore, the difference between experimental group 1 (*Mdn* = 27 %) and experimental group 2 (*Mdn* = 27 %) in the pretest was found as not significant *(Exact Mann-Whitney-U*($N_{\text{Experimental group 1}}$ = 14, $N_{\text{Experimental group 2}}$ = 16,) = 106.000, *z* = -0.251, *one-tailed p* = .406, *r* = 0.05), as well as the difference between experimental group 1 (*Mdn* = 27 %) and the control group (*Mdn* = 24 %) in the pretest (*Exact Mann-Whintey-U*($N_{\text{Experimental group 1}}$ = 14, $N_{\text{Control group}}$ = 16,) = 95.500, *z* = -0.690, *one-tailed p* = .251, *r* = 0.13) and the difference between experimental group 2 (*Mdn* = 27 %) and the control group (*Mdn* = 24 %) in the pretest (*Exact Mann-Whitney-U*($N_{\text{Experimental group 2}}$ = 16, $N_{\text{Control group}}$ = 16,) = 103.000, *z* = -0.948, *one-tailed p* = .177, *r* = 0.17). All analyses investigating the pretest and the proportion of correct responses can be found in Appendix K.

Significant difference was revealed in the posttest between the three groups (*Kruskal-Wallis-H (2)* = 7.423, *two-tailed p* = .024, *r* = 0.36). The difference between experimental group 1 (*Mdn* = 39 %) and experimental group 2 (*Mdn* = 34 %) was found significant and can be seen in **Table 2**. Another significant difference was revealed between experimental group 1 (*Mdn* = 39 %) and the control group (*Mdn* = 32 %) and can also be seen in **Table 2**. A not significant difference was found between experimental group 2 (*Mdn* = 34 %) and the control group (*Mdn* = 32 %), *Exact Mann-Whitney-U*($N_{\text{Experimental group 2}}$ = 16, $N_{\text{Control group}}$ = 16,) = 107.500, *z* = -0.777, *one-tailed p* = .224, *r* = 0.14. All analyses considering the posttest and the proportion of correct responses are documented in Appendix L.

Table 2

Exact Mann-Whitney-U-Tests regarding the Proportion of Correct Responses of Experimental Group 1, Experimental Group 2, and the Control Group in the Posttest

	Proportion of correct responses of experimental group 1		Proportion of correct responses of experimental group 2		Proportion of correct responses of the control group				
Variable	*Mdn*	*MD*	*Mdn*	*MD*	*Mdn*	*MD*	*U*	*p (one-tailed)*	Cohen's *r*
Experimental group 1 vs. experimental group 2 posttest	39 %	4.5 %	34 %	8.5 %	-	-	70.000	.039	0.32
Experimental group 1 vs. control group posttest	39 %	4.5 %	-	-	32 %	8 %	46.500	.002	0.50

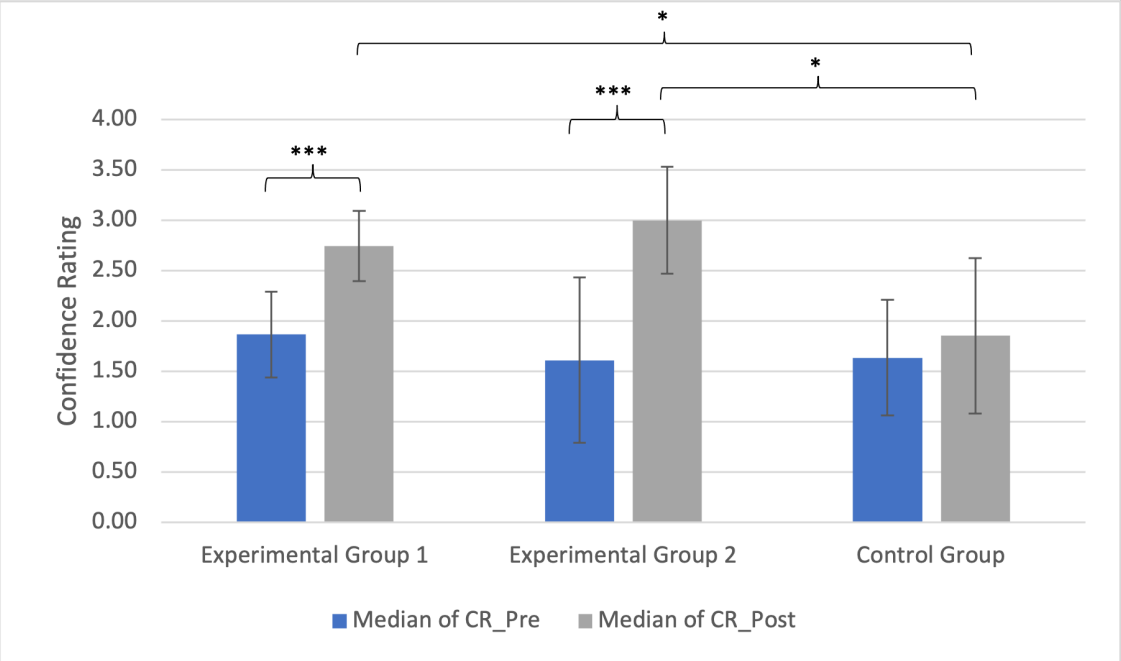
Note. Mdn = Median. MD = Mean deviation from the median. Experimental group 1 contains 14 participants, experimental group 2 contains 16 participants, and the control group contains 16 participants.

3.2 Confidence Rating

The hypothesis for the confidence rating investigated in this paper was that experimental group 1 would significantly increase their confidence in their given responses due to their training compared to experimental group 2 and the control group. Prior to the analyses of significance, a Kolmogorov–Smirnov test was conducted to investigate whether the variable was normally distributed. The confidence rating was not normally distributed in the pretest and therefore non-parametric methods were used for the analysis (experimental group 1; pretest: $p = .200$, posttest: $p = .200$ / experimental group 2; pretest: $p = .044$, posttest: $p = .078$ / control group; pretest: $p = .002$, posttest: $p = .200$). The analysis can be seen in Appendix M.

Figure 7

Overview of Confidence Rating



Note. The median for each group regarding the confidence rating (y-axis). Error bars show the mean deviation from the median. *p < .05, **p < .01, ***p < .001.

An overview is shown in **Figure 7** regarding the difference between experimental group 1, experimental group 2 and the control group considering the confidence rating for the pretest and posttest. To investigate the difference between the pre- and posttest for each group a Paired-Sample Wilcoxon Signed Rank Test was executed and visualized in Violin plots (Appendix N). Experimental group 1 revealed a significant difference between the pretest (*Mdn* = 1.87) and the posttest (*Mdn* = 2.74) and can be seen in **Table 3**. Additionally, experimental group 2, also shown in **Table 3**, also revealed a significant difference between the pretest (*Mdn* = 1.61) and the posttest (*Mdn* = 3.00). No significant difference could be found for the control group for the pretest (*Mdn* = 1.63) and posttest (*Mdn* = 1.85), $T = 73$, $z = -0.739$, *one-tailed* $p = .240$, $n = 16$, $r = 0.18$. The boxplot in **Figure 8** shows the differences in pretest and posttest regarding the confidence rating for all three groups.

Table 3

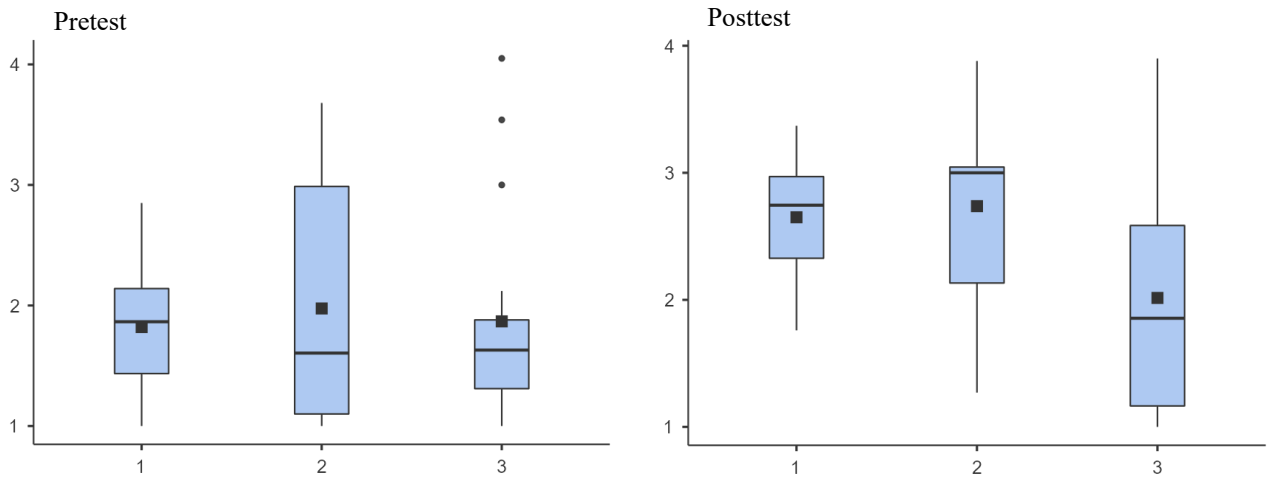
Paired-Sample Wilcoxon Signed Rank Test regarding the Pretest and Posttest of Experimental Group 1 and Experimental Group 2

	Confidence Rating of pretest		Confidence Rating of posttest					
Variable	*Mdn*	*MD*	*Mdn*	*MD*	*T*	*z*	*p (one-tailed)*	Cohen's *r*
Pre- and posttest for experimental group 1	1.87	0.43	2.74	0.35	105	-3.297	< .001	0.88
Pre- and posttest for experimental group 2	1.61	0.82	3.00	0.53	124	-2.896	.001	0.72

Note. Mdn = Median. MD = Mean deviation from the median. Experimental group 1 contains 14 participants and experimental group 2 contain 16 participants.

Figure 8

Boxplots visualizing Confidence Rating Pretest and Posttest



Note. Boxplots of the pretest (left) and posttest right) with confidence rating (y-axis). The line represents the median while ■ represents the mean. Mild outliers are shown by ● . 1 shows the results of experimental group 1, 2 shown the results of experimental group 2 and 3 shows the results of the control group.

The confidence rating in the pretest did not show any significant differences between the three groups (*Kruskal-Wallis-H (2)* = 0.249, *two-tailed p* = .883, *r* = 0.20). No significant difference was revealed between experimental group 1 (*Mdn* = 1.87) and experimental group 2 (*Mdn* = 1.61), *Exact Mann-Whitney-U*($N_{\text{Experimental group 1}}$ = 14, $N_{\text{Experimental group 2}}$ = 16,) = 108.500, *z* = -0.146, *one-tailed p* = .447, *r* = 0.03. Also, no significant difference between experimental group 1 (*Mdn* = 1.87) and the control group (*Mdn* = 1.63) was found (*Exact Mann-Whitney-U*($N_{\text{Experimental group 1}}$ = 14, $N_{\text{Control group}}$ = 16,) = 95.500, *z* = -0.687, *one-tailed p* = .252, *r* = 0.13) as well as between experimental group 2 (*Mdn* = 1.61) and the control group (*Mdn* = 1.63), *Exact Mann-Whitney-U*($N_{\text{Experimental group 2}}$ = 16, $N_{\text{Control group}}$ = 16,) = 127.000, *z* = -0.038, *one-tailed p* = .489, *r* = 0.01. All analysis investigating the confidence rating in the pretest can be seen in Appendix O.

Focusing on the posttest, a significant difference between the three groups was revealed (*Kruskal-Wallis-H (2)* = 6.849, *two-tailed p* = .033, *r* = 0.34). No significant difference between experimental group 1 (*Mdn* = 2.74) and experimental group 2 (*Mnd*= 3.00) was shown (*Exact Mann-Whitney-U*($N_{\text{Experimental group 1}}$ = 14, $N_{\text{Experimental group 2}}$ = 16,) = 97.000, *z* = -0.625, *one-tailed p* = .272, *r* = 0.11), but between experimental group 1 (*Mnd* = 2.74) and the control group (*Mnd* = 1.85) which can be seen in **Table 4**. Additionally, the difference between experimental group 2 (*Mdn* = 3.00) and the control group (*Mdn* = 1.85) was also revealed as significant also shown in **Table 4**. All analysis of the confidence rating in the posttest are documented in Appendix P.

Table 4

Exact Mann-Whitney-U-Tests regarding the Confidence Rating of Experimental Group 1, Experimental Group 2, and the Control Group in the Posttest

	Confidence rating of experimental group 1		Confidence rating of experimental group 2		Confidence rating of the control group				
Variable	*Mdn*	*MD*	*Mdn*	*MD*	*Mdn*	*MD*	*U*	*p (one-tailed)*	Cohen's *r*
Experimental group 1 vs. control group posttest	2.74	0.35	-	-	1.85	0.77	59.500	.014	0.40
Experimental group 2 vs. control group posttest	-	-	3.00	0.53	1.85	0.77	68.500	.012	0.40

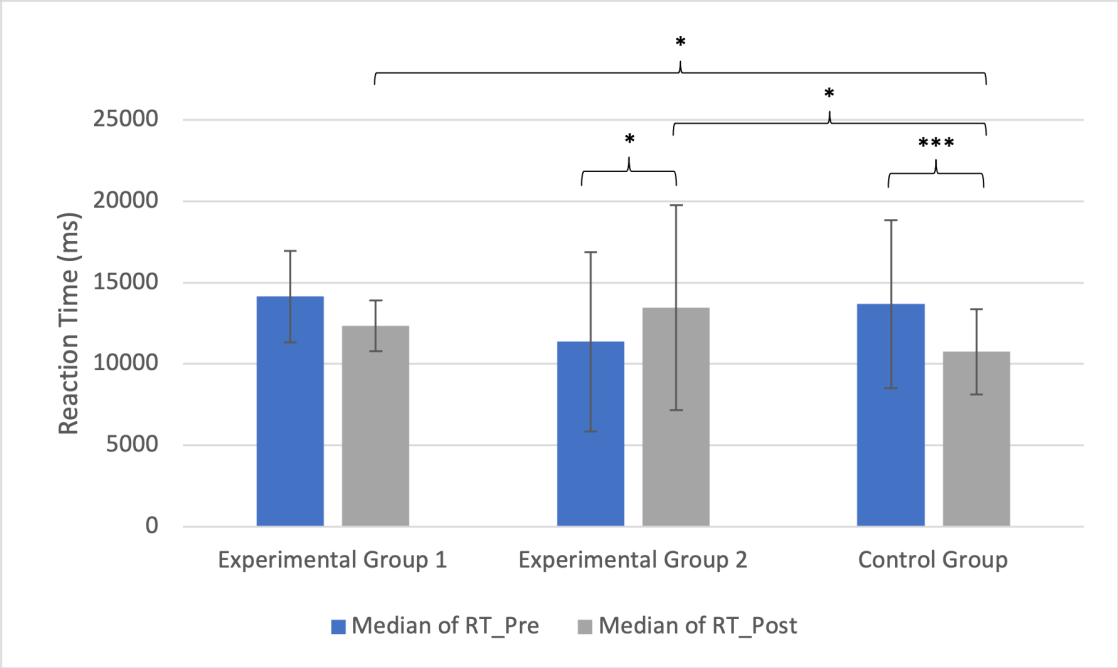
Note. Mdn = Median. MD = Mean deviation from the median. Experimental group 1 contains 14 participants, experimental group 2 contains 16 participants, and the control group contains 16 participants.

3.3 Reaction Time

The hypothesis for the reaction time was that experimental group 1 would be significantly faster in reaching a decision regarding the disease or abnormality on the ultrasound image by pressing on a response button than experimental group 2 and the control group due to the training. The M-UCAT measured the reaction time, and the unit of measurement was milliseconds. The reaction time describes the duration from the moment when the ultrasound image appears to the moment when the participant clicks on the chosen answer button. To start the analysis, a Kolmogorov–Smirnov test was executed to investigate if the variable is normally distributed. Reaction time was not normally distributed in the pre- and posttest, therefore non-parametric methods were used for the analysis (experimental group 1; pretest: $p = .200$, posttest: $p = .003$ / experimental group 2; pretest: $p = .010$, posttest: $p = .093$ / control group; pretest: $p = .006$, posttest: $p = .200$). The analysis can be seen in Appendix Q.

Figure 9

Overview of Reaction Time



Note. The median for each group regarding the reaction time in milliseconds (y-axis). Error bars show the mean deviation from the median. *p < .05, **p < .01, ***p < 001.

Figure 9 shows the difference between experimental group 1, experimental group 2 and the control group considering the reaction time for the pre- and posttest. Comparing the pretest (*Mdn* = 14153) and posttest (*Mdn* = 12341) using a Paired-Sample Wilcoxon Signed Rank Test of experimental group 1 revealed no significant difference ($T = 29$, $z = -1.475$, *one-tailed* $p = .077$, $n = 14$, $r = 0.39$). However, the difference in experimental group 2 between pretest (*Mdn* = 11384) and posttest (*Mdn* = 13454) was found significant and can be seen in **Table 5**. Another significant difference was revealed between the pretest (*Mdn* = 13695) and posttest (*Mdn* = 10753) in the control group which also can be seen in **Table 5**. **Figure 10** shows a Boxplot containing the results from the pretest and the posttest regarding the reaction time of all three

groups. All analysis of the Paired-Sample Wilcoxon Signed Rank Tests and the visualization with Violin plots can be seen in Appendix R.

Table 5

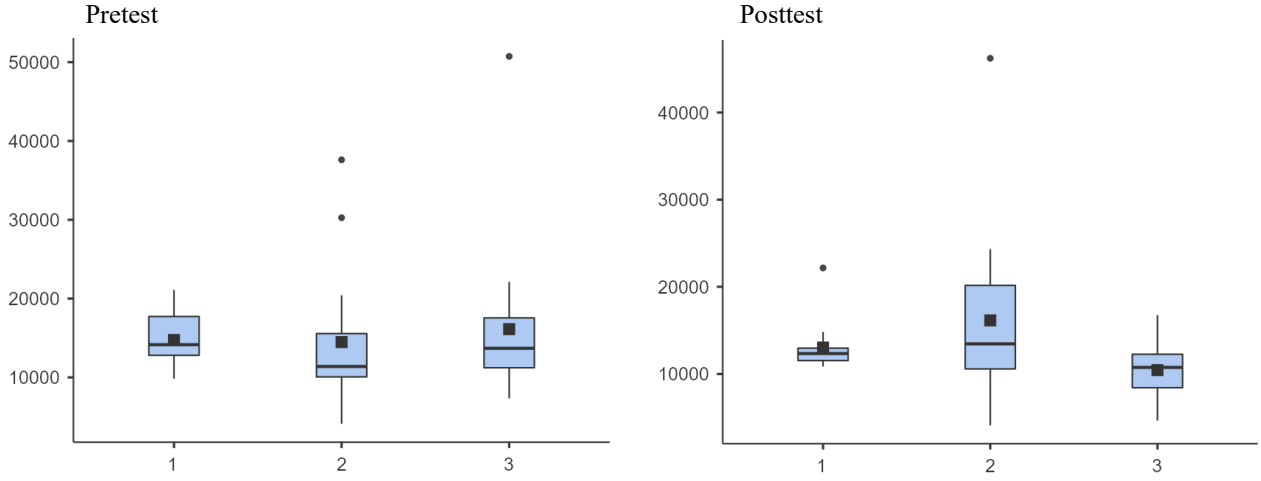
Paired-Sample Wilcoxon Signed Rank Test Regarding the Pretest and Posttest of Experimental Group 2 and the Control Group

	Reaction time of pretest		Reaction time of posttest					
Variable	*Mdn*	*MD*	*Mdn*	*MD*	*T*	*z*	*p (one-tailed)*	Cohen's *r*
Pre- and posttest for experimental group 2	11384	5506	13454	6294	103	-1.810	.037	0.45
Pre- and posttest for the control group	13695	5158	10752	2613	5	-3.258	< .001	0.81

Note. Mdn = Median. MD = Mean deviation from the median.

Figure 10

Boxplots visualizing Reaction Time Pretest and Posttest



Note. Boxplots of the pretest (left) and posttest (right) with reaction time in milliseconds (y-axis). The line represents the median while ■ represents the mean. Mild outliers are shown by●. 1 shows the results of experimental group 1, 2 shown the results of experimental group 2 and 3 shows the results of the control group.

Focusing on the pretest no significant difference could be found between the three groups, *Kruskal-Wallis-H (2)* = 1.266, *two-tailed p* = .531, *r* = 0.13. No significant difference could be found between experimental group 1 (*Mdn* = 14153) and experimental group 2 (*Mdn* = 11384), *Exact Mann-Whitney-U*($N_{\text{Experimental group 1}}$ = 14, $N_{\text{Experimental group 2}}$ = 16,) = 88.000, *z* = -0.998, *one-tailed p* = .167, *r* = 0.18. Also, between experimental group 1 (*Mdn* = 14153) and the control group (*Mdn* = 13695) no significant difference was revealed (*Exact Mann-Whitney-U*($N_{\text{Experimental group 1}}$ = 14, $N_{\text{Control group}}$ = 16,) = 103.000, *z* = -0.374, *one-tailed p* = .364, *r* = 0.07) as well as between experimental group 2 (*Mdn* = 11384) and the control group (*Mdn* = 13695), *Exact Mann-Whitney-U*($N_{\text{Experimental group 2}}$ = 16, $N_{\text{Control group}}$ = 16,) = 105.000, *z* = -0.867, *one-tailed p* = .201, *r* = 0.15. All analysis of the reaction time in the pretest can be found in Appendix S.

A not significant difference was revealed in the posttest between the three groups, *Kruskal-Wallis-H (2)* = 5.691, *two-tailed p* = .058, *r* = 0.30. The difference between experimental group 1 (*Mdn* = 12341) and experimental group 2 (*Mdn* = 13454) was found not significant, *Exact Mann-Whitney-U*($N_{\text{Experimental group 1}}$ = 14, $N_{\text{Experimental group 2}}$ = 16,) = 101.000, *z* = -0.457, *one-tailed p* = .334, *r* = 0.08. A significant difference was found between experimental group 1 (*Mdn* = 12341) and the control group (*Mdn* = 10752) and is shown in **Table 6**. Another significant difference was revealed between experimental group 2 (*Mdn* = 13454) and the control group (*Mdn* = 10751) and also can be seen in **Table 6**. All analysis regarding the reaction time of the posttest is documented in Appendix T.

Table 6

Exact Mann-Whitney-U-Tests regarding the Reaction Time of Experimental Group 1, Experimental Group 2, and the Control Group in the Posttest

	Reaction time experimental group 1		Reaction time experimental group 2		Reaction time control group				
Variable	*Mdn*	*MD*	*Mdn*	*MD*	*Mdn*	*MD*	*U*	*p (one-tailed)*	Cohen's *r*
Experimental group 1 vs. control group posttest	12341	1552	-	-	10752	2613	61.000	.017	0.39
Experimental group 2 vs. control group posttest	-	-	13454	6294	10752	2613	76.000	.026	0.35

Note. Mdn = Median. MD = Mean deviation from the median. Experimental group 1 contains 14 participants, experimental group 2 contains 16 participants, and the control group contains 16 participants.

3.4 Questionnaire

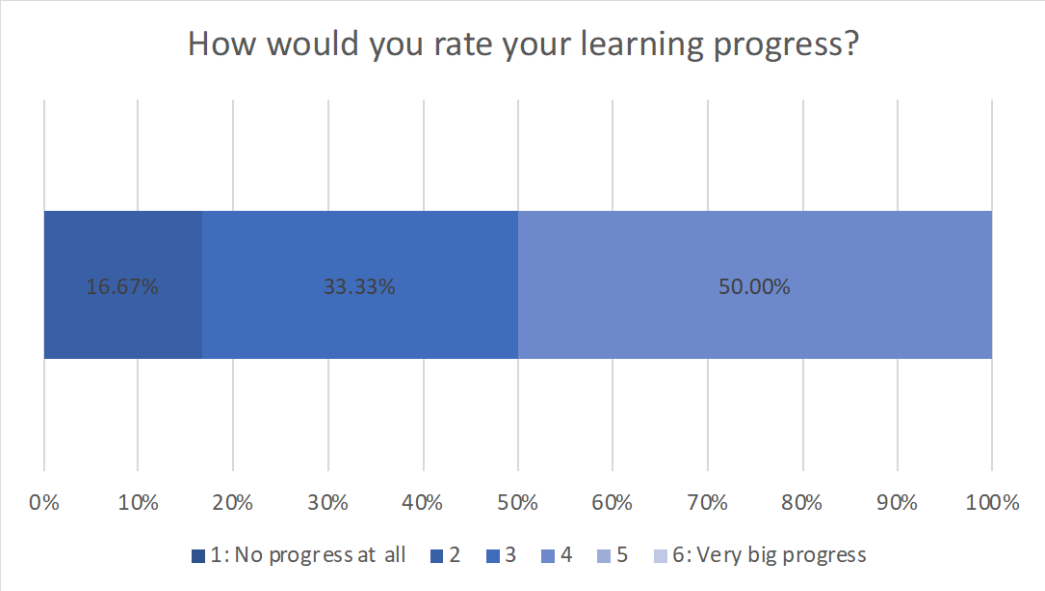
As described in chapter 2.3.1, the second questionnaire was only sent to experimental group 1 to give them the chance to state their opinion about the M-UTT, training regarding the detection performance in medical images in general, and how visual search is currently taught at their university. Six of the 14 participants in experimental group 1 answered the questions in the survey, leading to a response rate of 43 %. The results shown in the following paragraphs are also used to improve the M-UTT for further studies.

In general, the M-UTT was liked by the participants. 66.67 % of the participants chose the number 4 or 5 on a scale from 1 to 6, where 1 stated “Not liked at all” and 6 specified “Extremely well liked”. 16.67 % of the participants chose the number 2, and an equal amount chose the number 3 (Appendix U). Regarding the question about what they particularly liked particularly about the M-UTT, they responded mostly with the opportunity to repeat the wrongly declared images and that one could look up content during the training. They also mentioned the division into blocks of different topics regarding diseases or abnormalities of the gallbladder as helpful and the possibility that the M-UTT could be conducted at home (Appendix V). On the other hand, the participants also stated their opinion about their dislikes of the M-UTT. The majority of the answers were about the usability of the M-UTT, like the answer buttons or the limited response time. Also, the feedback for each image was not specific enough. The fact that images were repeatedly shown even though the participant assigned the image correctly was not liked by the participants. Functions the participants missed in the M-UTT were a block at the end, where the images from the previous blocks were mixed, as well as specific feedback after each image why the not correct answers could not be suitable for the specific image (Appendix w).

Regarding the individual perceived learning progress through the M-UTT, 50 % of the participants stated a medium opinion but on the positive side (number 4). 33.33 % also answered a medium opinion but more on the negative side (number 3). One person (16.67 %) did receive even less individual learning progress and chose number 2 on the scale from 1 “No progress at all” to 6 “Very big progress”, which can be seen in **Figure 11**.

Figure 11

Illustration of “How would you rate your learning progress?”



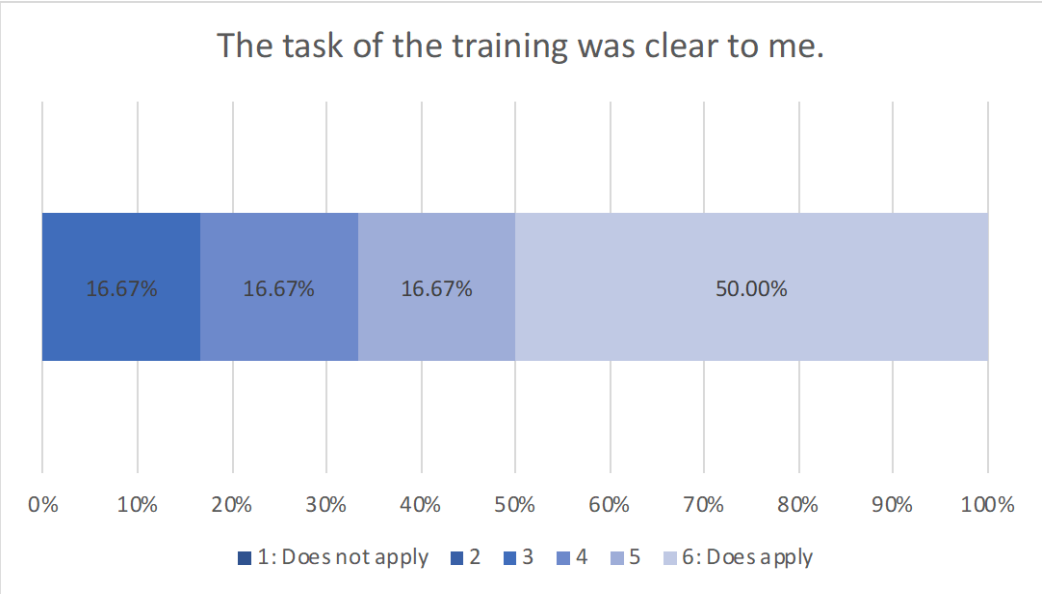
Note. n = 6

The participants generally liked the usability of the M-UTT. 50 % of the participants chose a medium answer, but on the positive side (number 4) on a scale from 1 “Does not apply” and 6 “Does apply” regarding the statement “The training is user-friendly”. In comparison, 16.67% chose answer option 3, and 33.33% answered with option 5. (Appendix X). Regarding the usefulness of the M-UTT, the participants were divided in their opinion. Towards the statement “I found the feedback during the training useful”, on a scale from 1 “Does not apply” to 6 “Does apply”, 33.33 % chose number 2 and 16.67 % chose number 3, which is more of a medium expression but on

the negative side of the scale. However, 33.33 % chose answer number 5, and 16.67 % selected number 6, expressing that the training was received as applicable (Appendix Y). Seen in **Figure 12** is the question with the statement "The task of the training was clear to me". 50 % of the participants expressed that the statement fully applies by choosing the highest available answer option. Answer numbers 5, 4, and 3 were each chosen once. When asked whether the participants had fun participating in the training with the M-UTT, the majority responded positively. 50 % chose to answer with option number 4, while 33.33 % answered option number 5, which equals 5 of 6 persons who chose the positive options on the scale from 1 "Does not apply" to 6 "Does apply" (Appendix Z).

Figure 12

Illustration of "The task of the training was clear to me."



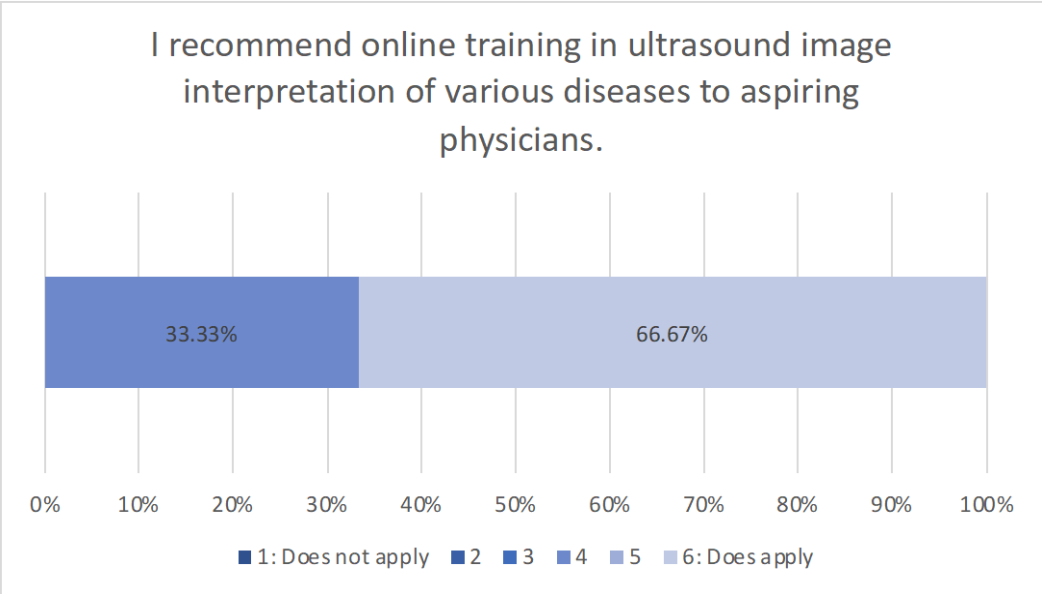
Note. n = 6

The last statement regarding the training focused on training in general, not specifically the M-UTT. All the participants responded positively to the statement whether they would recommend an online training for aspiring physicians. 33.33 %

chose answer option 4, while 4 of the participants (66.67 %) chose answer number 6 on the scale from 1 “Does not apply” to 6 “Does apply” (**Figure 13**).

Figure 13

Illustration of “I recommend online training in ultrasound image interpretation of various diseases to aspiring physicians.”



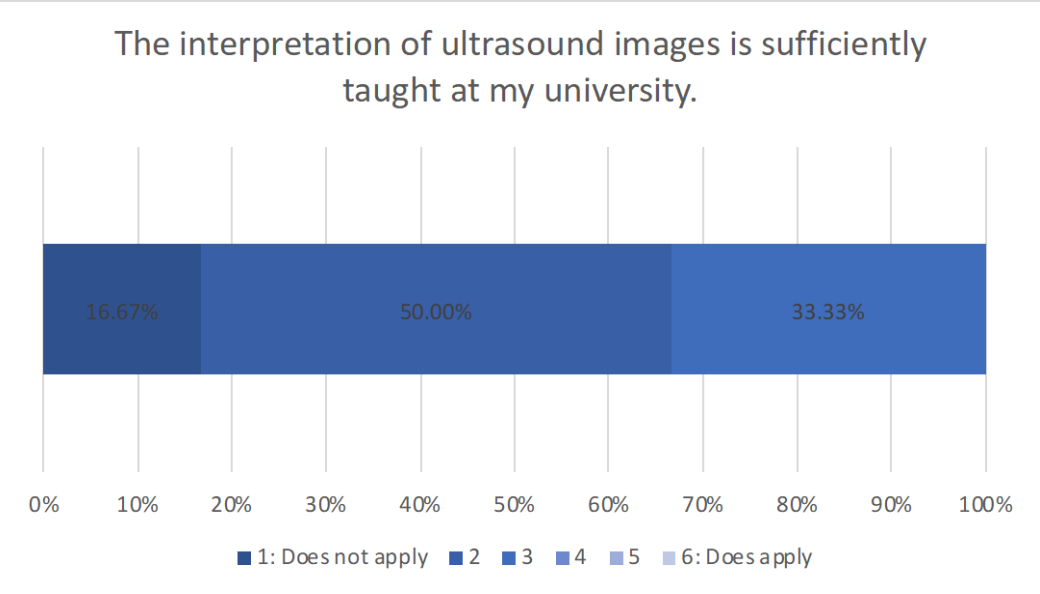
Note. n = 6

As mentioned in the beginning, the questionnaire also gave the participants the opportunity to express their opinion about the extent to which ultrasound interpretation is currently taught at their university. The participants responded negatively to the statement which defined the teaching of interpretation skills of ultrasound images as sufficient. All answers were on the negative side of the scale. Precisely one person chose answer option number 1, 3 persons chose answer option 2, and 2 persons number 2 on the scale from 1 “Does not apply” to 6 “Does apply” (**Figure 14**). They described the teaching as that the interpretation of ultrasound images is introduced in radiology courses but only superficial and the visit of an independent training in a skill lab where the medical students teach each other (peer to peer) (Appendix AA). However, the

participants agreed with the statement that an online training-type module in the curriculum of their university would make sense, as seen in **Figure 15**. All participants expressed their opinion on the positive side of the scale from 1 "Not at all useful" to 6 "Extremely useful" by choosing answer option number 5 or 6.

Figure 14

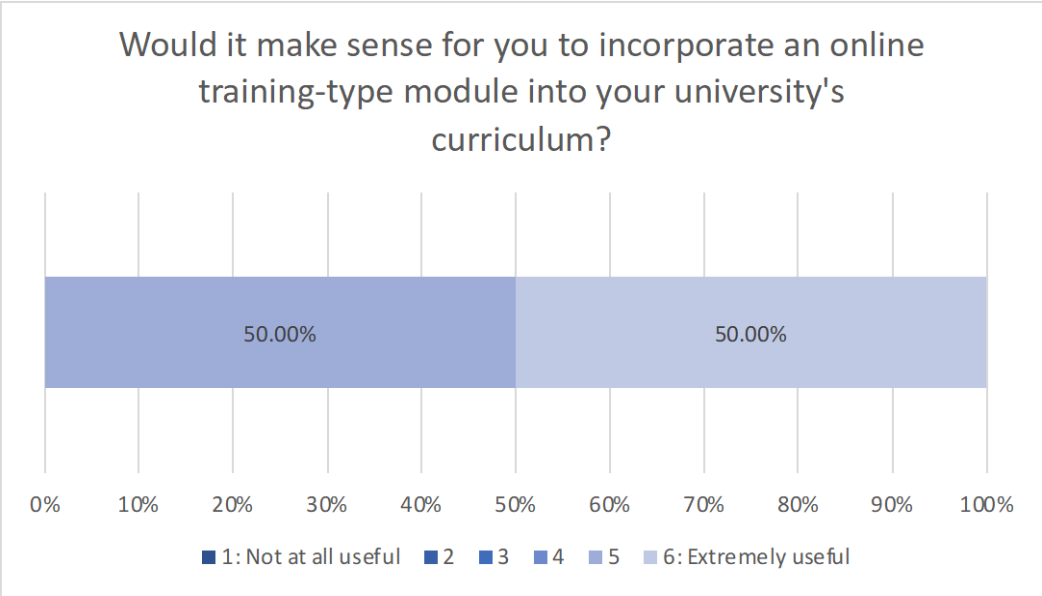
Illustration of "The interpretation of ultrasound images is sufficiently taught at my university."



Note. n = 6

Figure 15

Illustration of "Would it make sense for you to incorporate an online training-type module into your university's curriculum?"



Note. n = 6

To focus on their opinion about the idea of an online training-type module in their university's curriculum, one of the last questions in the questionnaire asked about in which year of study they would recommend the training module be introduced. Three participants recommended including the training module in the fourth study year, while the other three defined the range from the third to the sixth study year or specifically the fifth and the sixth study year (Appendix BB).

4 Discussion

The presented study focuses on the visual search task in medicine. Even for well-trained experts, the visual search task in medicine is still defined as complex (Wolfe, 2010). Due to the fact that medicine is considered a high-consequence field and false detection of medical images can come with financial and emotional costs, the importance of the visual search task in medicine is demonstrated (Carrigan et al., 2015;

See et al., 2017; Wolfe, 2010). Radiologists are experts in examining medical images, which have become an inherent part of modern healthcare but need comprehensive training and practice to reach their highly accurate performance level (Carrigan et al., 2015; Yang et al., 2002). A study using realistic VR simulation training could show significant improvements in medical students' skill levels (Gunn et al., 2018). The sonographic search task contains different aspects than the visual search task of radiologists. It is somewhat unique, being a continuous visual and cognitive task and inherits making decisions in real-time (Carrigan et al., 2015). Due to the gaining attention of POCUS, where attending physicians examine the patient directly at the bedside and its wide range of application areas in medicine, POCUS is focused on the presented study. (Moore & Copel, 2011; Osterwalder & Tercanli, 2018). Yamada, Minami, Soni, et al. (2018) could show that a one-day POCUS training is enough to significantly enhance the POCUS knowledge, image interpretation skills, and confidence in their skills of novice trainees and novice attending physicians. As mentioned in chapter 1, this study is based on the not published study of Lehmann and Michel (2020), where a similar testing and training system was used based on the work in aviation security (*CASRA*, n.d.). Declaring this training system as an active learning method due to the actively engaging of the participants with the software and the instant received feedback, the theory and previous research on active learning were considered in this study. To examine the theory of McDonald and Frank (2016), that the combination of passive and active learning scores even better results in an abstract concept learning task than active learning alone, supports the two different training sessions of experimental group 1 and experimental group 2.

Finally, several papers mentioned the need for a standardized training for POCUS (Mendiratta-Lala et al., 2010; Osterwalder & Tercanli, 2018; Yamada et al.,

2018). The presented study is an attempt at this request. It combines it with the beneficial outcomes of early exposure by using medical students as participants and the positive outcomes of combining passive with active learning methods (Branstetter et al., 2007; McDonald & Frank, 2016).

Before and after the training session, a pretest and posttest were conducted to investigate the effect of the respective training. Detection performance could not be analyzed by the Signal Detection Theory explained in chapter 2.4, so therefore the proportion of correct responses was used instead. A rating of the participant's confidence was also measured on the two test dates as well as the time until the participants reached a decision since POCUS is also used in emergency medicine, where time is a critical factor (Fatovich, 2002). To expose every participant to the same amount of knowledge beforehand about ultrasound, all participants listened to the first oral presentation provided by an expert in sonography and medical education. After, the first test (pretest, M-UCAT) was conducted. For the training session between the two tests, experimental groups 1 and 2 both listened to the passive learning content, the second oral presentation. After the second oral presentation, experimental group 1 additionally conducted the computer-based training (M-UTT). The control group did not engage in any kind of training. At last, the second test (M-UCAT) was executed by all participants. A questionnaire was sent to the affected participants to allow experimental group 1 to express their opinion about the M-UTT.

Following the results of this study will be discussed along with the research question and the three hypotheses.

4.1 Detection Performance

As introduced in chapter 1, the detection performance was conducted in this study. The hypothesis stated that the detection performance would significantly increase

through the combination of the passive training session (oral presentation) and the active training session (computer-based training).

Since a declaration of the terms of the Signal Detection Theory by Green & Swets (1966) for visual search in medicine is unclear, the proportion of correct responses is measured and analyzed instead of the detection performance. The results in chapter 3.1 show that experimental group 1, who engaged in the oral presentation as a passive training session and the computer-based training as an active training method, could significantly increase their proportion of correct responses. Not only were the results between the pretest and the posttest in experimental group 1 highly significant, but they also showed significantly better results in the proportion of correct responses in the posttest compared to experimental group 2 and the control group. Compared to the control group, the significant increase was somewhat expected because the control group did not conduct any training. In contrast, experimental group 1 engaged in a beneficial combination of passive and active learning, according to McDonald and Frank (2016). Interestingly, experimental group 1 responded significantly better than experimental group 2, who only conducted the passive learning session (oral presentation). As described in the literature, the active learning method also showed in this study a more promising outcome than the passive learning method seen in the significant increase of experimental group 1 in the posttest. Furthermore, one of the facts that describe passive learning methods is that the professor can deliver knowledge to a bigger number of students simultaneously (Wingfield & Black, 2005). If a computer-based training like the M-UTT used in this study were included in a medical curriculum, this description would also be valid for an active learning method. The professor could include the knowledge into the info sheets of the training where the students could actively engage with the medical images. The aspect of time regarding

the absorption of knowledge is discussed in a different setting. Dorestani (2005) state that it is an advantage of the passive learning method that a professor can convey a lot of knowledge in a short amount of time. Yamada et al. (2018) challenged the necessary training time for POCUS and could show in their study that one-day of training already showed significant results. In the presented study, experimental group 1 conducted the oral presentation of 29 minutes and the computer-based training, where they trained for 54 minutes on average. Due to the results shown in this study, participants who trained passively and actively for 1 hour and 23 minutes could significantly increase their proportion of correct responses compared to participants who did not conduct any kind of training. Furthermore, they also could increase their proportion of correct responses significantly compared to the participants who only conducted a 29-minute passive learning session within 54 minutes of active learning to the additional 29 minutes of passive learning. Hence, it is possible to significantly increase the proportion of correct responses for POCUS in less than a day. However, whether only a few hours of training is the right way to educate medical students is questionable, which is also shown in a response of a participant of de M-UTT who would have preferred several smaller training sessions compared to one long one. Finally, the fulfillment of the hypothesis stated for the detection performance remains unclear due to the challenging aspects of the Signal Detection Theory by Green and Swets (1966) in the medical field. However, the analyses of the proportion of correct responses show significant results, hence the hypothesis for an increase in correctly processing the ultrasound images will be accepted.

Through the positive findings in proportion of correct responses, the combination of the oral presentation and the computer-based training can be used to create a standardized training and a test to certify medical students who are interested in

POCUS. Furthermore, the standardized training would meet one of the measures stated by Osterwalder ad Tercanli (2018) but also could be included in the curriculum of medical universities, which was supported by the responses from the participants in the second questionnaire. The only critique of the training and test used in this study is the usability. Therefore, the response buttons' design, feedback changes, and other changes should be focused on when preparing for the standardized training and test and continuously updated.

4.2 Confidence Rating

This study's hypothesis regarding the confidence rating stated that the participants who conducted the combination training of passive and active learning methods would be significantly more confident than the other participants. Due to the findings reported in chapter 3.2, this hypothesis can be rightfully accepted.

Experimental group 1, which conducted the passive and active learning method (oral presentation and computer-based training), could increase the confidence in their response significantly from the pretest to the posttest as well as in the posttest compared to the control group. In combination with the finding from the proportion of correct responses discussed in chapter 4.1, it can be concluded that the combination of the passive and active learning methods not only increases the proportion of correct responses but also increases the confidence in their decisions. However, interesting are the responses in the second questionnaire of experimental group 1. The participants stated that when asked about their learning progress, they would rate it medium progress, even though they showed a significant increase of confidence between the pretest and posttest. Regarding the progress of their confidence, it can be assumed that the progress was not enough for the participants. In the posttest, they scored a median of 2.74, which is still relatively low compared to the highest possible response opportunity

of five but nonetheless is a significant increase from the pretest median of 1.87. By always deciding on a rather low confidence level, the participant may transfer this perception of confidence into the second questionnaire. Another explanation could be that the learning progress is not noticeable to the participants and, therefore, should be made visible in the training software or after the conducted tests.

Focusing on experimental group 2, who conducted the passive learning method through the oral presentation, the findings reported in chapter 3.2 are also significant. Experimental group 2 increased their confidence between the pretest and the posttest significantly and also compared to the control group in the posttest. Considering the findings from the proportion of correct responses discussed in chapter 4.1 where experimental group 2 could not increase the proportion of correct responses between the pretest and the posttest and also not compared to the control group in the posttest but increased their confidence level significantly. In other words, the participants of experimental group 2 were more confident in the posttest than in the pretest, even though the increase in their proportion of correct responses was not significant. This finding could lead to the theory of overconfidence described by Friedman et al. (2005), where participants conducting a little bit of training overestimate their abilities. It would also fit into the description of the "beginner's bubble" introduced by Sanchez and Dunning (2018). Because the participants in this study also were beginners with non to little knowledge about ultrasound images as they stated that the interpretation of ultrasound images is not sufficiently taught at their university in the second questionnaire. As described by Sanchez and Dunning (2018), overconfidence deflates after further exposure to training and knowledge, which cannot be observed in this presented study but would be an interesting aspect for further research. Especially regarding the development of POCUS because confidence about the declared diagnosis

is expected from physicians (Croskerry & Norman, 2008). Additionally, stating a diagnosis as an attending physician, even though the visual search task in medicine is demanding as described by Wolfe (2010), should come with a certain amount of confidence to then benefit from the advantages that POCUS brings, such as low cost or reduced time to establish a diagnosis (Moore & Copel, 2011; Osterwalder & Tercanli, 2018). Experimental group 1 may also be overconfident in their decisions, but it is difficult to observe because a group with participants with prior knowledge and experience is missing in this study.

The findings for the control group are not surprising. The participants could not significantly increase their confidence level between the pretest and posttest and showed lower confidence than the other two experimental groups in the posttest. Furthermore, since they did not engage in any kind of training, they also could not increase their proportion of correct responses discussed in chapter 4.1. Therefore, they have not more confidence in their responses because they lack an increase in their knowledge which could have acted as a basis for also increasing their confidence in their responses.

4.3 Reaction Time

The reaction time, which means the time from when the image was presented to the time when the participants reached a decision by pressing an answer button, was measured in this study. In chapter 1, the hypothesis was described that the group that conducted the combination of passive and active learning methods (oral presentation and computer-based training) would decrease their reaction time significantly and, therefore, be faster in reaching a decision than the other two groups.

As the results reported in chapter 3.3 show, experimental group 1 did decrease their reaction time from the pretest to the posttest, but this difference was not significant. Therefore, a speed-accuracy-trade-off cannot be observed because

experimental group 1 increases their proportion of correct responses and shows lower reaction time after the passive and active training sessions, even though the difference is insignificant. As mentioned in Bogacz et al. (2010) and Franks et al. (2003), a speed-accuracy-trade-off describes the trade between speed and accuracy and would therefore mean that participants are faster but less accurate or more accurate but slower. An explanation of why the difference between pretest and posttest in experimental group 1 is not significant can be that the duration of the training which was considered as rather short by the participants in the second questionnaire. The importance of time for POCUS is clearly stated since POCUS is also used in emergency situations (Fatovich, 2002; Moore & Copel, 2011; Osterwalder & Tercanli, 2018).

Experimental group 2 showed a significant difference in reaction time between the pretest and the posttest. Surprisingly, the group took significantly more time to reach a decision and choose one of the displayed answer buttons after they concluded the oral presentation as a passive learning method. Combined with the findings discussed in chapter 4.1, where experimental group 2 increased their proportion of correct responses slightly and not significantly, a trend toward a speed-accuracy-trade-off could be assumed. The passive learning method would then influence the participants towards emphasizing accuracy over speed. Hence the participants need more time to investigate an ultrasound image.

The control group who did not engage in any training showed surprising results in the measure of reaction time. Their reaction time significantly decreased from the pretest to the posttest and showed significant differences compared to the reaction times of experimental groups 1 and 2 in the posttest. Since the control group did not show significant differences in the proportion of correct responses, discussed in chapter 4.1, and no significant differences in the confidence in their responses, discussed in chapter

4.2, this finding in the reaction time could point towards a trend of speed-accuracy-trade-off. Speed was clearly emphasized in the posttest in the control group, but the participants did not significantly decrease in their proportion of correct responses, which should appear if a speed-accuracy-trade-off occurs as described in the literature (Bogacz et al., 2010; Franks et al., 2003). The lack of training could explain why the control group focused on speed in the posttest. The participants could not increase their knowledge and could therefore only control their response time in the posttest.

4.4 Limitations

Different aspects limit the presented study. The possibility of participating in this study and executing the tests and pieces of training at home or in a place of the participants' choosing can be seen as critical. Even though this opportunity was liked by the participants, as stated in the second questionnaire by experimental group 1, the circumstances such as the lighting of the room, amount of background noise, or other aspects which could influence the participants during the task were not controlled by the study team and are therefore unknown. However, if a test and training were included in medical universities' curricula, this aspect would probably remain true. Another task that the study team could not control was the oral presentations. Even though the participants were instructed to listen to the oral presentations first before requiring access to the next step in the study by e-mail, there was no possibility to check if every participant listened to the whole oral presentations.

As mentioned in chapters 2.3.3 and 2.3.4, the test and training software used in this study were initially developed for airport security, are still in use today, and are based on the Signal Detection Theory (see *CASRA*, n.d.; Green & Swets, 1966). As explained in chapter 2.4, this theory could not be used in the presented study. Further and more detailed analysis could be made if the aspects (e.g., correct rejection, false

alarm, miss, hit) were possible to include in this study. Therefore, the lack of inclusion of the Signal Detection Theory limits the presented study.

A limitation of the analysis is also that the data was not normally distributed and, therefore, must be analyzed with non-parametric methods. Although a test of significance was still possible and showed interesting results, further analysis, such as the interaction between two variables, was not possible.

4.5 Further Research

Despite the limitation, the presented study could show that computer-based training (active learning method) combined with a passive learning method (oral presentation) increases the proportion of correct responses of ultrasound images of changes in the gallbladder in medical students. Further research should focus on connecting the Signal Detection Theory with medical imaging to benefit from the aspects of the theory as it can be seen in, for example, airport security. Furthermore, knowledge about what aspects impede the investigation of medical images, what correct rejection and false alarm mean in medicine, or the actual ratio of images with an abnormality or disease compared to images without can help to understand the visual search task in medicine even more and could also be used by CASRA to develop a test and training system.

As mentioned in chapter 1, POCUS is used in different medical fields. Therefore, research regarding images of different body parts respectively organs should be pursued. Another interesting area of research would be the change in the medium of education. In this study, a computer-based training was used as an active learning method. However, other methods would be imaginable, like group discussions or VR, which is already used in radiology studies described in chapter 1.

Further research specifically related to the finding of this study could also be pursued. For example, regarding the reaction time, it would be interesting if significant findings could be found in experimental groups with further training. Also, the aspect of motivation could play an interesting role, especially in a group that does not engage in training. Another result was the trend toward overconfidence in experimental group 2. As seen in the literature cited in chapter 1, let us assume that also a possibility of overconfidence could be found in experimental group 1 if compared with experienced physicians. Further research in this area would bring clarity to the aspect of overconfidence and the amount and kind of training in POCUS.

5 Conclusion

According to the results revealed in this study, combining an active and passive learning method, specifically here the combination of an oral presentation and a computer-based training, can be used successfully to train medical students in their proportion of correct responses while examining ultrasound images of the gallbladder. Especially since the group who only conducted the passive learning method did not increase their proportion of correct responses significantly highlights the ability of the computer-based training when before important information is lectured in the oral presentation. Therefore, computer-based training is a learning method that can be used to educate medical students in visual search in medicine. Furthermore, it is a possibility to fulfill the need for standardized training for POCUS required from the medical community. It could also lead to a certifying process using a test format like it was presented in this study. Therefore, this study also supports the spread of the use of POCUS, even though concerns from the medical community are justified, but the necessity of suitable training for POCUS is highlighted. An example of training for POCUS is presented in this study.

Further research should focus on the possibility to standardizes a computer-based training for POCUS. For example, aspects of overconfidence, speed-accuracy-trade-off, and the effect of a combination of active and passive learning methods should be further examined and noted in the standardization process. Another focus could be to include other human body areas in the computer-based training and studies, where also attending physicians take part as compared to medical students.

A new field of opportunity opens for CASRA, which is already active in train and certifying aviation security screeners. Visual search in medicine, standardized training for medical students or attending physicians, and a possible certifying process could present an exciting area for the company. This study shows that the already used training and testing system can be altered and adapted to the medical area. However, the system's usability requires further alternation to fit medical personnel's needs regarding POCUS training. Also, the expansion to other body areas and the recommendation for further research should be pursued by CASRA. To successfully enter the field of visual search in medicine, the connection to medical experts, who already supported the presented study, needs to be maintained, and connections ultimately need to expand to hospitals and universities with a medical department.

6 References

Alba, G. A., Kelmenson, D. A., Noble, V. E., Murray, A. F., & Currier, P. F. (2013). Faculty staff-guided versus self-guided ultrasound training for internal medicine residents. *Medical Education*, *47*(11), 1099–1108. https://doi.org/10.1111/medu.12259

Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, *103*(4), 718–735. https://doi.org/10.1037/a0029395

Armor, D. A., & Taylor, S. E. (1998). Situated Optimism: Specific Outcome Expectancies and Self-Regulation. In *Advances in Experimental Social Psychology* (Vol. 30, pp. 309–379). Elsevier. https://doi.org/10.1016/S0065-2601(08)60386-X

Barnett, S. B. (2002). Routine ultrasound scanning in first trimester: What are the risks? *Seminars in Ultrasound, CT and MRI*, *23*(5), 387–391. https://doi.org/10.1016/S0887-2171(02)90009-0

Benek-Rivera, J., & Mathews, V. E. (2004). Active Learning with Jeopardy: Students Ask the Questions. *Journal of Management Education*, *28*(1), 104–118. https://doi.org/10.1177/1052562903252637

Berlin, L. (2005). Errors of Omission. *American Journal of Roentgenology*, *185*(6), 1416–1421. https://doi.org/10.2214/AJR.05.0838

Bogacz, R., Hu, P. T., Holmes, P. J., & Cohen, J. D. (2010). Do humans produce the speed–accuracy trade-off that maximizes reward rate? *Quarterly Journal of Experimental Psychology*, *63*(5), 863–891. https://doi.org/10.1080/17470210903091643

Bonwell, C., & Eison, J. A. (1991). Active Learning: Creating Excitement in the Classroom. ERIC Digest. *ERIC Clearinghouse on Higher Education,*

Washington, DC.; George Washington Univ., Washington, DC.

Branstetter, B. F., Faix, L. E., Humphrey, A. L., & Schumann, J. B. (2007). Preclinical Medical Student Training in Radiology: The Effect of Early Exposure. *American Journal of Roentgenology*, *188*(1), W9–W14. https://doi.org/10.2214/AJR.05.2139

Brenner, D. J., & Hall, E. J. (2007). Computed Tomography—An Increasing Source of Radiation Exposure. *The New England Journal of Medicine*, *357*(22), 2277–2284.

Bridge, P., Gunn, T., Kastanis, L., Pack, D., Rowntree, P., Starkey, D., Mahoney, G., Berry, C., Braithwaite, V., & Wilson-Stewart, K. (2014). The development and evaluation of a medical imaging training immersive environment. *Journal of Medical Radiation Sciences*, *61*(3), 159–165. https://doi.org/10.1002/jmrs.60

Carrigan, A. J., Brennan, P. C., Pietrzyk, M., Clarke, J., & Chekaluk, E. (2015). A 'snapshot' of the visual search behaviours of medical sonographers. *Australasian Journal of Ultrasound in Medicine*, *18*(2), 70–77. https://doi.org/10.1002/j.2205-0140.2015.tb00045.x

CASRA. (n.d.). CASRA. Retrieved March 19, 2022, from https://www.casra.ch

Cohen, J. C. (2004). *Instituting improvement in medical education. 13*(11), 2 (publication of the AAMC).

Croskerry, P., & Norman, G. (2008). Overconfidence in Clinical Decision Making. *The American Journal of Medicine*, *121*(5), S24–S29. https://doi.org/10.1016/j.amjmed.2008.02.001

Dorestani, A. (2005). Is Interactive/Active Learning Superior to Traditional Lecturing in Economics Courses? *Humanomics*, *21*(1), 1–20. https://doi.org/10.1108/eb018897

Dunning, D., & Griffin, D. W. (1990). The Overconfidence Effect in Social Prediction. *Journal of Personality and Social Psychology*, *58*(4), 568–581. http://dx.doi.org/10.1037/0022-3514 .58.4.568

Ebert-May, D., Brewer, C., & Allred, S. (1997). Innovation in Large Lectures: Teaching for Active Learning. *BioScience*, *47*(9), 601–607. https://doi.org/10.2307/1313166

Fatovich, D. M. (2002). Recent developments: Emergency medicine. *BMJ*, *324*(7343), 958–962. https://doi.org/10.1136/bmj.324.7343.958

Franks, N. R., Dornhaus, A., Fitzsimmons, J. P., & Stevens, M. (2003). Speed versus accuracy in collective decision making. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1532), 2457–2463. https://doi.org/10.1098/rspb.2003.2527

Friedman, C. P., Gatti, G. G., Franz, T. M., Murphy, G. C., Wolf, F. M., Heckerling, P. S., Fine, P. L., Miller, T. M., & Elstein, A. S. (2005). Do physicians know when their diagnoses are correct?: Implications for decision support and error reduction. *Journal of General Internal Medicine*, *20*(4), 334–339. https://doi.org/10.1111/j.1525-1497.2005.30145.x

Graffam, B. (2007). Active learning in medical education: Strategies for beginning implementation. *Medical Teacher*, *29*(1), 38–42. https://doi.org/10.1080/01421590601176398

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley & Sons.

Greenstein, Y. Y., Littauer, R., Narasimhan, M., Mayo, P. H., & Koenig, S. J. (2017). Effectiveness of a Critical Care Ultrasonography Course. *Chest*, *151*(1), 34–40. https://doi.org/10.1016/j.chest.2016.08.1465

Gunn, T., Jones, L., Bridge, P., Rowntree, P., & Nissen, L. (2018). The use of virtual reality simulation to improve technical skill in the undergraduate medical imaging student. *Interactive Learning Environments*, *26*(5), 613–620. https://doi.org/10.1080/10494820.2017.1374981

Hossfeld, B. (2020). *Ultraschall – das neue Stethoskop des Notaufnahmearztes?* News Paper. https://news-papers.eu/?p=12172

Hurst, J. W. (2004). The Overlecturing and Underteaching of Clinical Medicine. *Archives of Internal Medicine*, *164*(15), 1605. https://doi.org/10.1001/archinte.164.15.1605

Karkhanis, V., & Joshi, J. (2012). Pleural effusion: Diagnosis, treatment, and management. *Open Access Emergency Medicine*, 31. https://doi.org/10.2147/OAEM.S29942

Kundel, H. L., Nodine, C. F., & Carmody, D. (1978). *Visual Scanning, Pattern Recognition and Decision-making in Pulmonary Nodule Detection*. *13*(3), 175–181. https://doi.org/10.1097/00004424-197805000-00001

Lehmann, M., & Michel, S. (2020). *Improving the quality of radiological findings through computer-based training*.

Maw, A., Jalali, C., Jannat-Khah, D., Gudi, K., Logio, L., Evans, A., Anderson, S., & Smith, J. (2016). Faculty development in point of care ultrasound for internists. *Medical Education Online*, *21*(1), 33287. https://doi.org/10.3402/meo.v21.33287

McDonald, K., & Frank, M. (2016). *When does passive learning improve the effectiveness of active learning*.

McGraw, A. P., Mellers, B. A., & Ritov, I. (2004). The affective costs of overconfidence. *Journal of Behavioral Decision Making*, *17*(4), 281–295.

https://doi.org/10.1002/bdm.472

Mendiratta-Lala, M., Williams, T., de Quadros, N., Bonnett, J., & Mendiratta, V. (2010). The Use of a Simulation Center to Improve Resident Proficiency in Performing Ultrasound-Guided Procedures. *Academic Radiology*, *17*(4), 535–540. https://doi.org/10.1016/j.acra.2009.11.010

Michel, N., Cater, J. J., & Varela, O. (2009). Active versus passive teaching styles: An empirical study of student learning outcomes. *Human Resource Development Quarterly*, *20*(4), 397–418. https://doi.org/10.1002/hrdq.20025

Miner, F. C., Das, H., & Gale, J. (1984). An Investigation of the Relative Effectiveness of Three Diverse Teaching Methodologies. *Journal of Management Education*, *9*(2), 49–59. https://doi.org/10.1177/105256298400900207

Moore, C. L., & Copel, J. A. (2011). Point-of-Care Ultrasonography. *The New England Journal of Medicine*, *364*(8), 749–757.

Osterwalder, J., & Tercanli, S. (2018). POCUS – Chance or Risk? *Ultraschall in Der Medizin - European Journal of Ultrasound*, *39*(06), 606–609. https://doi.org/10.1055/a-0720-8864

Sanchez, C., & Dunning, D. (2018). Overconfidence among beginners: Is a little learning a dangerous thing? *Journal of Personality and Social Psychology*, *114*(1), 10–28. https://doi.org/10.1037/pspa0000102

Sarason, Y., & Banbury, C. (2004). Active Learning Facilitated by Using a Game-Show Format or Who Doesn't Want to be a Millionaire? *Journal of Management Education*, *28*(4), 509–518. https://doi.org/10.1177/1052562903260808

See, J. E., Drury, C. G., Speed, A., Williams, A., & Khalandi, N. (2017). The Role of Visual Inspection in the 21 st Century. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 262–266.

https://doi.org/10.1177/1541931213601548

Van Eynde, D. F., & Spencer, R. W. (1988). Lecture Versus Experiential Learning: Their Differential Effects On Long-Term Memory. *Journal of Management Education*, *12*(4), 52–58. https://doi.org/10.1177/105256298801200404

Wingfield, S. S., & Black, G. S. (2005). Active Versus Passive Course Designs: The Impact on Student Outcomes. *Journal of Education for Business*, *81*(2), 119–123. https://doi.org/10.3200/JOEB.81.2.119-128

Wolfe, J. M. (2010). Visual search. *Current Biology*, *20*(8), 346–349. https://doi.org/doi:10.1016/j.cub.2010.02.016

Yamada, T., Minami, T., Soni, N. J., Hiraoka, E., Takahashi, H., Okubo, T., & Sato, J. (2018). Skills acquisition for novice learners after a point-of-care ultrasound course: Does clinical rank matter? *BMC Medical Education*, *18*(1), 202. https://doi.org/10.1186/s12909-018-1310-3

Yang, G.-Z., Dempere-Marco, L., Hu, X.-P., & Rowe, A. (2002). Visual search: Psychophysical models and practical applications. *Image and Vision Computing*, *20*(4), 291–305. https://doi.org/10.1016/S0262-8856(02)00022-7

7 List of Figures

8 List of Tables